Linking health survey data with health insurance data: benefits and opportunities for public health research

Finaba BERETE

Thesis submitted in fulfilment of the requirements for the doctoral degree in Public Health Sciences

Academic year 2023-2024

Promotors: Prof. dr. Olivier BRUYÈRE
Prof. dr. Herman VAN OYEN

LIÈGE université

LIÈGE université Médecine

Jury Members

Prof. dr. Olivier BRUYÈRE (Promotor)	University of Liège
Prof. dr. Herman VAN OYEN (Promotor)	Sciensano, Ghent University
Prof. dr. Pierre GILLET (President)	University hospital of Liège, University of Liège
Prof. dr. Olivier ETHGEN (Secretary)	University of Liège
dr. Johan VAN DER HEYDEN	Sciensano
Prof. dr. Katrien VANTHOMME (External member)	Ghent University
Prof. dr. Romana HANEEF (External member)	Santé Publique France

Finaba Berete

Health Information

Epidemiology and public health

Sciensano

Juliette Wytsmanstraat 14, 1050 Brussels

Belgium

finaba.berete@sciensano.be

Abstract

Population-based surveys such as national health interview surveys (HISs) are essential tools to collect information on health status, use of health care and health determinants in the general population. However, the validity of self-reported information through surveys is a concern due to the associated selection and reporting biases. In addition to these validity issues (as a result of selection and reporting bias), HISs are also facing other challenges due to the increasing need of researchers for more comprehensive data to answer complex research questions. Increasing the number of questions in HIS may result in a high workload for interviewers and a significant burden on respondents. This would lead to dropouts resulting in missing data and lower response rates, which both affect data quality.

Data linkage has become a popular approach in the secondary use of existing data. It is being increasingly used in health services research, longitudinal studies, disease surveillance and health policy. Data linkage has been found to be a useful and efficient approach for obtaining more complete data without increasing the length of the questionnaires. In addition, data linkage can play a crucial role to get further insights about the validity of self-reported information.

In Belgium, the Belgian health interview survey (BHIS) and the Belgian compulsory health insurance (BCHI) data are important population-based data sources. Linking the two data sources (HISlink) allows on the one hand a validation of some of the survey data, and results on the other hand in a richer database which offers new research opportunities useful for public health authorities.

Based on the use case of the HISlink, the overarching aim of this thesis was to investigate the potential benefits and opportunities of linking health survey data with health insurance data for public health research.

To explore the fundamental concepts of data linkage, a literature review was undertaken to cover the following questions: What is data linkage? What are commonly the types of linked data? What methods have been used to link data? What are the challenges and the legal issues? How to assess the quality of linked data?

Then, the following two research questions were examined:

1. To what extent can linked data be used to assess data validity?

This question was explored for three topics: self-reported mammography uptake, chronic diseases, and polypharmacy.

2. To what extent can linked data be used to respond to policy-relevant questions which cannot be addressed with each of the sources separately?

This question was explored for two policy-relevant research questions: what are predictors of nursing home admission among the older population? What is the mediating effect of health literacy in the relationship between socioeconomic status and health outcomes.

Although the objectives of this thesis are quite broad, offering a wide range of research possibilities, only a limited number of topics were selected. This selection was guided by the relevance of the topics for public health (relevance for the commissioner of the linkage, relevance with respect to societal challenges), and the feasibility in the relation to the information available in both databases.

What is data linkage? What are commonly the types of linked data? What methods have been used to link data? What are the challenges and the legal issues? How to assess the quality of linked data?

Data linkage brings together information that relates to the same individual, family, place or event from different data sources. Varying data sources within the context of health public research can be linked together, including surveys data (e.g., health interview surveys, health examination surveys, social surveys) and administrative data (e.g., health insurance claims data, hospital discharge data, prescription drugs data, electronic medical records, diseases-specific registries).

Two approaches have been identified to undertake a linkage: the deterministic (rule-based) methods where an exact match on all linking variables is required, and the probabilistic (score-based) methods where a match weight (score) is assigned to represent the likelihood that two records belong to the same individual. The content and quality of the data sources to be linked play an important role in the choice of the linkage methods. The deterministic methods are simplest and best suited to 'perfect' data where there are unique personal identifiers or highly discriminating linkage keys.

The probabilistic methods are the most commonly used approaches because perfect data are rare, however these methods are also more complex to implement.

Privacy and confidentiality issues remain the key concerns in data linkage. Linkage errors pose the greatest threat to the quality of linked data and ultimately may lead to information bias and selection bias. Care must therefore be taken to assess the quality of the linkage in order to provide reliable results. Several methods are proposed to assess the quality of the linked data including standard metrics (e.g., match rate, recall, precision, etc.) or more elaborated approaches (e.g., comparison with gold standard, sensitivity analysis, comparison linked vs. unlinked data, quality control check, etc.). Researchers must validate the linked data before undertaking any analysis.

To what extent can linked data be used to assess data validity

The potential of linked data to be used for validity analysis was presented in three papers.

The first paper entitled *"Validity of self-reported mammography uptake in the Belgian health interview survey: selection and reporting bias "* (chapter 4) focused on mammography uptake and examined the validity of BHIS 2013 information on this topic, using BCHI data as the gold standard. This case study revealed considerable differences in the prevalence of mammography uptake among women aged 50–69 years in the BHIS source (75.5%), compared to the BCHI source (69.8%) and the random sample of the BCHI (64.1%). The validity of BHIS information regarding mammography uptake was significantly affected by both selection bias (assessed through the comparison of BCHI based estimates for the participants included in the BHIS sample and the same estimates in the random sample of the BCHI) and reporting bias (assessed through a comparison of BHIS-based estimates with BCHI-based estimates within the same sample). The relative size of selection bias and reporting bias was 9% and 8%, respectively. The reporting bias, which turned out to be mainly related to 'telescoping' (i.e. remembering that an event occurred more recently than it actually did) was unequally distributed across population subgroups.

The second paper entitled *"Comparing administrative and survey data for ascertaining chronic disease prevalence* (chapter 4*)* compared BHIS and BCHI data for ascertaining the prevalence of a selection of chronic diseases. Good agreement for

diabetes, cardiovascular diseases (including hypertension), Parkinson's disease and thyroid disorders (kappa between 0.63 and 0.77), moderate agreement for epilepsy (kappa = 0.46) and poor agreement for chronic obstructive pulmonary disease and asthma (kappa = 0.35). The agreement was influenced by individual socio-demographic characteristics and health status, although their effects varied from one chronic disease to another.

The third paper entitled "*Assessing polypharmacy in the older population: comparison of a self-reported and prescription based method*" (chapter 4) compared the two data sources for estimating the prevalence of polypharmacy and assessed their complementarity. The study highlighted that within the population aged 65 years and older, self-reported and prescription based polypharmacy prevalence estimates were respectively 27% and 32%. Overall agreement was moderate, but better in men (kappa = 0.60) than in women (kappa = 0.45). Determinants of moderate polypharmacy did not vary substantially by data source.

In summary, findings from these three validity studies suggested that the data collected as part of a health interview survey differs from that found in administrative data sources, and this varies according to the specific topics under consideration and the characteristics of the survey participants. Although both data sources have their advantages, relying solely on one would result in a poor estimate of the indicator in question. For this reason, objective data should be combined with survey data wherever possible.

To what extent can linked data be used to respond to policy-relevant questions which cannot be addressed with each of the sources separately?

This question was addressed in two papers.

The first paper entitled "Predictors of nursing home admission in the older population in Belgium: a longitudinal follow-up of health interview survey participants" (chapter 5) showed how the linkage of BHIS data with longitudinal BCHI data can be used to estimate the cumulative risk of nursing home admission among the older population of 65+ years (BCHI data) and its predictors (BHIS data) in Belgium. The cumulative risk of nursing home admission was 1.4%, 5.7% and 13.1% at, respectively 1 year, 3 years and 5 years of follow-up. The factors predicting nursing home admission are

multifactorial: a higher age, living situation (social supports), history of falls, urinary incontinence, physical chronic conditions and mental disorders such as Alzheimer's disease, appeared as strong predictors of nursing home admission. The findings suggested that preventing falls, managing urinary incontinence at home and providing appropriate and timely management of limitations, depression and Alzheimer's disease would delay the onset of nursing home admission.

The second paper entitled "Does health literacy mediate the relationship between socioeconomic status and health related outcomes in the Belgian adult population?" (chapter 6) presented a case study investigating the mediating effects of health literacy on the relationship between socioeconomic status as measured by education, income and a selected health and health related outcomes in varying domains: 1) health behaviour (physical activity, diet, alcohol and tobacco consumption), 2) health status (perceived health status, mental health), 3) use of medicines (purchase of antibiotics), and 4) use of preventive care (preventive dental care, influenza vaccination, breast cancer screening). Health literacy was found to partially mediate the association between socioeconomic status and physical activity, type of diet, alcohol and tobacco consumption, perceived health status, mental health and preventive dental care. The mediating effect of health literacy accounted for 2.5% to 15.4% of the total effect, suggesting that improving health literacy might reduce socioeconomic disparities in these domains. There was no significant mediating effect of health literacy in the pathway through which socioeconomic status affects the purchase of antibiotics, influenza vaccination and breast cancer screening.

These two case studies showed that in public health, to answer certain research questions the use of multiple data sources is required. In such cases, data linkage is a powerful tool for obtaining a richer database from which to carry out the necessary analyses. Both of these studies could not have been be carried out with one database. Specifically, while some information can only be extracted from administrative data sources (for example, date of entry into the nursing home), other information can only be obtained through health surveys (such as health status, health behaviour, social support). Furthermore, thanks to the linked data, researchers can choose to combine information from the two sources in order to obtain a more accurate indicator or to choose the source of the information according to the confidence placed in the source for this information.

Conclusions and recommendations

This thesis demonstrates that data linkage brings significant added value to public health research. It makes it possible to assess the validity of data sources and to answer policy-relevant research questions that cannot be answered using separate tools. However, linking survey data to administrative data is challenging because of privacy considerations and time consuming. The work undertaken in the context of this thesis have a number of implications. Caution should be exercise when using survey and administrative data separately to produce policy-relevant indicators. As much as possible, the linkage of both data sources should be used. From a public health perspective, policy makers should continue investing in data linkages; taking up initiatives to work towards a better balance between the right to privacy of respondents and society's right to evidence-based information to improve health; facilitating access and reuse of data including data linkage through tools like the European health data space or national data linkage hubs. Researchers should consider improving the communication with the surveys participants, so there is more willingness to give a consent for linkage.

Although based on Belgian data and in Belgian specific context, we believe that this study has a much broader implications and could be useful to researchers who plan to link health survey data with health administrative data for their respective projects.

Résumé

Les enquêtes populationnelles, telles que les enquêtes nationales de santé par interview, sont des outils essentiels pour collecter des informations sur l'état de santé, l'utilisation des soins de santé et les déterminants de la santé au sein de la population générale. Toutefois, la validité des informations autodéclarées dans le cadre des enquêtes peut être mise à mal en raison de biais de sélection et de déclaration qui y sont associés. Outre les problèmes de validité que ces biais entraînent, les enquêtes de santé par interview sont également confrontées à d'autres défis en raison du besoin croissant des chercheurs de disposer de données plus complètes pour répondre à des questions de recherche complexes. L'augmentation du nombre de questions dans les enquêtes peut entraîner une charge de travail élevée pour les enquêteurs et un fardeau important pour les répondants. Il en résulterait des abandons qui se traduiraient par des données manquantes et des taux de réponse plus faibles, ce qui affecterait la qualité des données.

Le couplage de données issues des enquêtes avec d'autres sources d'informations individuelles est devenu une approche populaire dans l'utilisation secondaire de données existantes. Elle est de plus en plus utilisée dans la recherche sur les services de santé, les études longitudinales, la surveillance des maladies et la politique de santé. Le couplage de données s'est avéré être une approche utile et efficace pour obtenir des données plus complètes sans augmenter la longueur des questionnaires. En outre, le couplage de données peut jouer un rôle crucial pour obtenir des informations supplémentaires sur la validité des informations autodéclarées.

En Belgique, les données de l'enquête de santé belge (BHIS) et de l'assurance maladie obligatoire belge (BCHI) sont d'importantes sources de données au niveau de la population. Le couplage des deux sources de données (HISlink) permet d'une part de valider certaines données de l'enquête, et d'autre part d'obtenir une base de données plus riche qui offre de nouvelles possibilités de recherche utiles aux autorités de santé publique.

Sur la base d'études de cas HISlink, l'objectif principal de cette thèse était d'étudier les avantages potentiels de coupler les données des enquêtes de santé aux données

de l'assurance maladie, et les perspectives qui sont offertes pour la recherche en santé publique.

Pour explorer les concepts fondamentaux du couplage des données, une revue de la littérature a été entreprise afin de répondre aux questions suivantes : Qu'est-ce que le couplage de données ? Quels sont les types de données couplées ? Quelles sont les méthodes utilisées pour coupler les données ? Quels sont les défis et les questions juridiques liés au couplage de données ? Comment évaluer la qualité des données couplées ?

Ensuite, les deux questions de recherche suivantes ont été examinées :

1. Dans quelle mesure les données couplées peuvent-elles être utilisées pour évaluer la validité des informations provenant de l'une ou de l'autre source ?

Cette question a été étudiée pour trois sujets : la mammographie autodéclarée, les maladies chroniques et la polypharmacie.

2. Dans quelle mesure les données couplées peuvent-elles être utilisées pour répondre à des questions stratégiques ne pouvant être traitées avec l'une ou l'autre source prise séparément ?

Cette question a été étudiée à partir de deux questions de recherche pertinentes pour les politiques de santé : quels sont les facteurs prédictifs de l'admission en maison de repos au sein de la population âgée ? Quel est l'effet médiateur de la littératie en santé dans la relation entre le statut socio-économique et les résultats en matière de santé ?

Bien que les objectifs de cette thèse soient assez vastes et offrent un large éventail de possibilités de recherche, seul un nombre limité de sujets a été sélectionné. Cette sélection a été guidée par la pertinence des sujets pour la santé publique (pertinence pour le commanditaire du couplage, pertinence par rapport aux défis sociétaux), et la faisabilité par rapport aux informations disponibles dans les deux bases de données.

Qu'est-ce que le couplage de données ? Quels sont les types de données couplées ? Quelles sont les méthodes utilisées pour coupler les données ? Quels sont les défis et les questions juridiques rencontrés ? Comment évaluer la qualité des données couplées ?

De manière générale, le couplage de données permet de rassembler des informations relatives à la même personne, à la même famille, au même lieu ou au même événement provenant de différentes sources de données. Dans le contexte de la recherche en santé publique, différentes sources de données peuvent être couplées, notamment les données d'enquêtes (par exemple, les enquêtes de santé par interview, les enquêtes de santé par examen, les enquêtes sociales) et les données administratives (par exemple, les données d'assurance maladie, les données sur les sorties d'hôpital, les données sur les médicaments délivrés sur ordonnance, les dossiers médicaux électroniques, les registres spécifiques à certaines maladies).

Deux approches ont été identifiées dans la littérature pour entreprendre un couplage : les méthodes déterministes (basées sur des règles), qui exigent une correspondance exacte pour toutes les variables de couplage, et les méthodes probabilistes (basées sur des scores), qui attribuent un poids (score) à la correspondance pour représenter la probabilité que deux enregistrements appartiennent au même individu. Le contenu et la qualité des sources de données à coupler jouent un rôle important dans le choix des méthodes de couplage. Les méthodes déterministes sont les plus simples et les mieux adaptées aux données "parfaites" lorsqu'il existe des identifiants personnels uniques ou des clés de couplage très discriminantes. Les méthodes probabilistes sont les plus couramment utilisées car les données parfaites sont rares, mais elles sont également plus complexes à mettre en œuvre.

Les questions de protection de la vie privée et de confidentialité des données restent les principales préoccupations en matière de couplage. Par ailleurs, les erreurs de couplage constituent la plus grande menace pour la qualité des données couplées et peuvent en fin de compte entraîner un biais d'information et un biais de sélection. Il faut donc veiller à évaluer la qualité du couplage afin de fournir des résultats fiables. Plusieurs méthodes sont proposées pour évaluer la qualité des données couplées, y compris des mesures standard (par exemple, le taux de correspondance, le rappel, la précision, etc.) ou des approches plus élaborées (par exemple, la comparaison avec l'étalon-or, l'analyse de sensibilité, la comparaison entre les données couplées

et les données non couplées, la vérification du contrôle de la qualité, etc.) Les chercheurs doivent valider les données couplées avant d'entreprendre toute analyse.

Dans quelle mesure les données couplées peuvent-elles être utilisées pour évaluer la validité des informations ?

Le potentiel des données couplées pour l'analyse de la validité de celles-ci a été présenté dans trois articles.

Le premier article intitulé *"Validity of self-reported mammography uptake in the Belgian health interview survey: selection and reporting bias"* (chapitre 4) portait sur le recours à la mammographie et examinait la validité des informations auto-rapportées provenant du BHIS 2013 sur ce sujet, en utilisant les données du BCHI comme étalon-or. Cette étude de cas a révélé des différences considérables dans la prévalence du recours à la mammographie chez les femmes âgées de 50 à 69 ans dans la source BHIS (75,5 %), par rapport à la source BCHI (69,8 %) et à l'échantillon aléatoire du BCHI (64,1 %). La validité des informations du BHIS concernant le recours à la mammographie a été significativement affectée par le biais de sélection (évalué par la comparaison des estimations basées sur le BCHI pour les participants inclus dans l'échantillon du BHIS et les mêmes estimations dans l'échantillon aléatoire du BCHI) et par le biais de déclaration (évalué par la comparaison des estimations basées sur le BHIS avec les estimations basées sur le BCHI au sein du même échantillon). L'importance relative du biais de sélection et du biais de déclaration était respectivement de 9 % et de 8 %. Le biais de déclaration, qui s'est avéré être principalement lié au "télescopage" (c'est-à-dire le fait de se souvenir qu'un événement s'est produit plus récemment qu'il ne s'est réellement produit), était inégalement réparti entre les sous-groupes de la population.

Le deuxième article intitulé *"Comparing administrative and survey data for ascertaining chronic disease prevalence"* (chapitre 4) a comparé les données du BHIS et du BCHI pour déterminer la prévalence d'une sélection de maladies chroniques dans la population générale. La concordance était bonne pour le diabète, les maladies cardiovasculaires (y compris l'hypertension), la maladie de Parkinson et les troubles thyroïdiens (kappa compris entre 0,63 et 0,77), modérée pour l'épilepsie (kappa = 0,46) et médiocre pour la bronchopneumopathie chronique obstructive et l'asthme (kappa = 0,35). La concordance a été influencée par les caractéristiques

sociodémographiques individuelles et l'état de santé, bien que leurs effets variaient d'une maladie chronique à l'autre.

Le troisième article intitulé *"Assessing polypharmacy in the older population: comparison of a self-reported and prescription based method"* (chapitre 4) a comparé les deux sources de données pour estimer la prévalence de la polypharmacie et a évalué leur complémentarité. L'étude a montré que, dans la population âgée de 65 ans et plus, les estimations de la prévalence de la polypharmacie autodéclarée et basée sur les ordonnances étaient respectivement de 27 % et 32 %. La concordance globale était modérée, mais meilleure chez les hommes (kappa = 0,60) que chez les femmes (kappa = 0,45). Les déterminants de la polypharmacie modérée ne variaient pas sensiblement selon la source de données.

En résumé, les résultats de ces trois études de validité suggèrent que les données collectées dans le cadre d'une enquête de santé par interview diffèrent de celles trouvées dans les sources de données administratives, et ce, en fonction des thèmes spécifiques considérés et des caractéristiques des participants à l'enquête. Bien que les deux sources de données aient leurs avantages, se baser uniquement sur l'une d'entre elles aboutirait à une mauvaise estimation de l'indicateur en question. C'est pourquoi les données objectives doivent être combinées avec les données d'enquête dans la mesure du possible.

Dans quelle mesure les données couplées peuvent-elles être utilisées pour répondre à des questions d'ordre politique qui ne peuvent être traitées avec chacune des sources séparément ?

Cette question a été abordée dans deux articles.

Le premier article intitulé "Predictors of nursing home admission in the older population in Belgium: a longitudinal follow-up of health interview survey participants" (chapitre 5) montre comment le couplage des données du BHIS avec les données longitudinales du BCHI peut être utilisé pour estimer le risque cumulé d'admission en maison de repos parmi la population âgée de 65 ans ou plus (données BCHI), et les prédicteurs d'admission en maison de repos en Belgique (données BHIS). Le risque cumulé d'admission en maison de repos était de 1,4 %, 5,7 % et 13,1 % à respectivement 1 an, 3 ans et 5 ans de suivi. Les facteurs prédictifs de l'admission en maison de repos sont multifactoriels : un âge élevé, la situation de vie (soutien social),

les antécédents de chutes, l'incontinence urinaire, les maladies chroniques physiques et les troubles mentaux tels que la maladie d'Alzheimer, sont apparus comme des facteurs prédictifs importants de l'admission en maison de repos. Les résultats suggèrent que la prévention des chutes, la gestion de l'incontinence urinaire à domicile et la prise en charge appropriée et opportune des limitations, de la dépression et de la maladie d'Alzheimer permettraient de retarder l'admission en maison de repos.

Le deuxième article intitulé "Does health literacy mediate the relationship between socioeconomic status and health related outcomes in the Belgian adult population?" (chapitre 6) présentait un cas étudiant les effets médiateurs de la littératie en santé sur la relation entre le statut socio-économique, mesuré par l'éducation, le revenu et une sélection de résultats en matière de santé dans différents domaines : 1) le comportement en matière de santé (activité physique, type d'alimentation, consommation d'alcool et de tabac), 2) l'état de santé (la santé perçue, la santé mentale), 3) l'utilisation de médicaments (achat d'antibiotiques) et 4) le recours aux soins préventifs (soins dentaires préventifs, vaccination contre la grippe, dépistage du cancer du sein). La littératie en santé s'est avérée être un médiateur partiel de l'association entre le statut socio-économique et l'activité physique, le type d'alimentation, la consommation d'alcool et de tabac, la santé perçue, la santé mentale et les soins dentaires préventifs. L'effet médiateur de la littératie en santé représentait de 2,5 % à 15,4 % de l'effet total, ce qui suggère que l'amélioration de la littératie en santé pourrait réduire les disparités socioéconomiques dans ces domaines. Il n'y a pas eu d'effet médiateur significatif de la littératie en santé dans la voie par laquelle le statut socio-économique affecte l'achat d'antibiotiques, la vaccination contre la grippe et le dépistage du cancer du sein.

Ces deux exemples montrent qu'en santé publique, l'utilisation de sources de données multiples est nécessaire pour répondre à certaines questions de recherche. Dans ce cas, le couplage de données est un outil puissant qui permet d'obtenir une base de données plus riche à partir de laquelle il est possible d'effectuer les analyses nécessaires. Ces deux études n'auraient pas pu être réalisées avec une seule base de données. En effet, si certaines informations ne peuvent être extraites que de sources de données administratives (par exemple, la date d'entrée dans la maison de repos), d'autres ne peuvent être obtenues que par le biais d'enquêtes de santé (telles

que l'état de santé perçue, le comportement en matière de santé, le soutien social). De plus, grâce aux données liées, les chercheurs peuvent choisir de combiner les informations des deux sources afin d'obtenir un indicateur plus précis ou de choisir la source de l'information la plus appropriée en fonction de la confiance accordée à la celle-ci pour cette information.

Conclusions et recommandations

Cette thèse démontre que le couplage de données apporte une valeur ajoutée significative à la recherche en santé publique. Elle permet d'évaluer la validité des sources de données et de répondre à des questions de recherche pertinentes pour la politique, auxquelles il est impossible de répondre à l'aide d'outils distincts. Cependant, la réalisation de couplage entre les données d'enquête et les données administratives est difficile en raison de considérations liées à la protection de la vie privée et prend du temps. Les travaux entrepris dans le cadre de cette thèse ont un certain nombre d'implications. Il convient d'être prudent lors de l'utilisation séparée des données d'enquête et des données administratives pour produire des indicateurs pertinents pour les politiques. Dans la mesure du possible, il convient d'utiliser le couplage entre les deux sources de données. Du point de vue de la santé publique, les décideurs politiques devraient continuer à investir dans les couplages de données, à prendre des initiatives pour trouver un meilleur équilibre entre le droit à la vie privée des personnes interrogées et le droit de la société à disposer d'informations fondées sur des données probantes pour améliorer la santé, à faciliter l'accès et la réutilisation des données, y compris le couplage de données, grâce à des outils tels que l'espace européen des données de santé ou les centres nationaux de couplage de données. Les chercheurs devraient envisager d'améliorer la communication avec les participants aux enquêtes, afin qu'ils soient plus enclins à consentir au couplage des données.

Bien que basée sur des données belges et dans le contexte spécifique de la Belgique, nous pensons que cette étude a une implication beaucoup plus large et pourrait être utile aux chercheurs qui prévoient de coupler les données d'enquêtes de santé aux données administratives de santé pour leurs projets respectifs.

Contents

Abbreviations

A	
ADL	Activities of Daily Living
ATC	Anatomical Therapeutic Chemical classification
B	
BCHI	Belgian Compulsory Health Insurance
BHIS	Belgian Health Interview Survey
C	
CAPI	Computer Assisted Personal Interviewing
CASI	Computer Aided Self Interviewing
CBSS	Crossroads Bank for Social Security
CCHS	Canadian Community Health Survey
CD	Chronic Disease
CELINE / IRCEL	Belgian Interregional Environment Agency (Cellule Interrégionale de l'Environnement / Intergewestelijke Cel voor het Leefmilieu)
CHMS	Canadian Health Measures Survey
CI	Confidence interval
CLSA	Canadian Longitudinal Study on Aging
CMS	Centers for Medicare & Medicaid Services
CNAV	National Old-Age Insurance Fund
CNK	National Code Number
CONOR	Cohort of Norway
COPD	Chronic Obstructive Pulmonary Disease
CSVD	Cerebral Small Vessel Disease
CVD	Cardiovascular disease
D	
DAD	Discharge Abstract Database
DDD	Daily Defined Dose
DPARD	Dutch Pediatric and Adult Registry of Diabetes
DPIA	Data Protection Impact Assessment
DPoRT	Diabetes Population Risk Tool
DPV	Diabetes-Patienten-Verlaufsdokumentation

DWH	Data warehouse
E	
EC	Ethics committee
ECDC	European Centre for Disease Prevention and Control
EDI	Early Development Instrument
EHDS	European Health Data Space
EHIS	European Health Interview Survey
EHIS-PAQ	Physical Activity Questionnaire developed by European Health Interview Survey
EM	Expectation Maximization
EMR	Electronic Medical Record
EPS	Échantillon Permantent(e) Steekproef
EU	European Union
F	
FAIR	Findable, Accessible, Interoperable and Reusable
FAMHP	Federal Agency for Medicines and Health Products
FPSHFCSE	Federal Public Services Public Health, Food Chain Safety and Environment
G	
GALI	Global Activity Limitations Indicator
GDPR	General Data Protection Regulation
GKV	Gesetzliche Krankenversicherung (German Statutory health insurance)
GP	General Practitioner
H	
HDA	Health Data Agency
HES	Hospital Episodes Statistics
HH	Household
HICN	Health Insurance Claim Number
HIS	Health Interview Survey
HISlink	Linkage between Belgian Health Interview Survey data and Belgian Compulsory Health Insurance data
HL	Health Literacy
HLS-EU-Q6	European Health Literacy Survey Questionnaire 6-items

HMDS	Hospital Morbidity Data System
HR	Hazard Ratio
I	
IAB	Integrated Employment Biographies
ICD	International Classification of Diseases
ICES	Institute for Clinical and Evaluative Sciences
ID	Individual Identification number
IDI	Integrated Data Infrastructure
IMA	InterMutualistic Agency
IQR	Interquartile range
IRB	Institutional Review Board
ISC	Information security committee
ISCED	International Standard Classification of Education
K	
KCE	Belgian Health Care Knowledge Centre
L	
LTC	Long Term Care
M	
MAF	Maximum A Facturer (Maximum billing)
MEHM	Minimum European Health Module
MHD	Minimum Hospital Data
MS	Member State
MSIS	Norwegian Surveillance System for Communicable Diseases
N	
NA	Non Available
NCHS	National Center for Health Statistics
NCI	National Cancer Institute
NH	Nursing Home
NHA	Nursing Home Admission
NHANES	National Health and Nutrition Examination Survey
NHI	National Health Index
NHIS	National Health Interview Survey

NHS	National Health Service
NHS-CR	National Health Service Central Register
NIC	National Intermutualist College
NIHDI	National Institute for Health and Disability Insurance
NPV	Negative Predictive Value
NR	National Register
NRN	Nationa Register Number
NZ	New Zealand
O	
ODB	Ontario Drug Benefit
OECD	Organisation for Economic Co-operation and Development
OHIP	Ontario Health Insurance Plan
OR	Odds Ratio
OTC	Over-the-Counter
P	
PAPI	Paper Assisted Personal Interviewing
PCG	Pharmacy Cost Groups
PII	Prior Informed Imputation
PPV	Positive Predictive Value
PSU	Primary Sample Unit
Q	
QPP	Quantity Per Package
R	
RAMQ	Regie de l'assurance maladie du Quebec
RDQ	Research, Development and Quality promotion
RN	Random Number
RRR	Report-to-Record Ratio
RWD	Real-word data
S	
SAS	Statistical Analysis System
SCRA	Small Cell Risk Analysis
SD	Standard Deviation
SE	Socioeconomic

SEER	Surveillance, Epidemiology and End Results
SES	Socioeconomic status
SHeS	Scottish Health Survey
SILC	Statistics on Income and Living Conditions
SLK	Statistical Linkage Key
SNDS	Système National de Données de Santé (The French National Healthcare Data System)
SNIIRAM	National inter-scheme health insurance information system
SRH	Self-rated health
SSN	Social Security Number
SSU	Secondary Sample Unit
STROBE	STrengthening the Reporting of OBservational studies in Epidemiology.
SVM	Support Vector Machines
T	
TILDA	The Irish Longitudinal Study on Ageing
TSU	Tertiary Sample Unit
TTP	Trusted Third Party
U	
UK	United Kingdom
UN	United Nations
US	United States
USA	United States of America
UZG	Universitair Ziekenhuis Gent
V	
VPN	Virtual Private Network
W	
WADLS	Western Australia Data Linkage System
WHO	World Health Organization

CHAPTER 1. GENERAL INTRODUCTION

1.1. BACKGROUND

Data linkage (or record linkage) is a method that brings together information that relates to the same individual, family, place or event from different data sources (1– 3) and is used to produce comprehensive data in a cost-effective way. Complex research questions require information from different data sources (integration of data), especially when the researcher does not have access to a rich database. A new and repeated primary data collection which covers all the dimensions that need to be considered is not only costly in terms of resources and a burden on the respondents, but is especially cost-inefficient where there is a possibility of linking with existing data. Therefore, researchers regularly opt for pooling independent data sources to obtain more comprehensive data in a cost-effective way. Internationally, data linkage is a common and widely accepted practice in public health for research addressing the use of health services, longitudinal studies, disease surveillance (4–7), and especially to leverage existing data. Linkages are increasingly used to generate evidence to inform policy and to guide health-service planning (3). In recent years, the secondary use of existing data has increased thanks to improved access arrangements, and data linkage has become one of the most cost-effective ways of supporting research in public health and epidemiology (8–11).

Different types of data can be brought together: one can either link diverse routine administrative data, or link survey data (preferably repeated) cross-sectional and longitudinal data collections to administrative data. Administrative data can be linked to health (e.g. hospital discharge data, sentinel networks, disease registries) but also to health determinants (socioeconomic status (SES), demographics, environment, etc.). In countries where administrative data linkage is well established (e.g. in the UK, Australia, Canada, Nordic countries), routinely linked administrative data sources are increasingly being used for public health research purposes (12–16), either by creating cohort studies (e.g. the Nordic registry-based cohort studies (14), the mother-baby cohort in England (17), the Danish open dynamic cohort (18) and the Melbourne Injecting Drug User Cohort Study (19)), or for specific study purposes (20,21). For instance, using a linkage of multiple administrative data sources (administrative workers' compensation claims data, universal health insurance data, state hospital and emergency department data, and social welfare data), Lane et al. (2021) were

able to estimate the impact of benefit cessation (income replacement cessation after 5 years) on healthcare service use in the UK (21). Moreover, the practice of linking survey data, from either cross-sectional (4,22) or longitudinal studies (4,23) with administrative data is often done to supplement survey information. In this context, the linkage can be project-based (ad hoc data linkage) or routine-based (systematic data linkage). Ad hoc data linkage is undertaken to support just one or a limited number of research projects, while systematic data linkage is undertaken on a proactive and regular basis for a population, with a view to supporting an indefinite number of future (and as yet undefined) health-related research projects.

Although the backbone of many (systematic) linkages relate to administrative data (i.e. registry-based linkage), this thesis will focus solely on the linkage of survey and administrative data and on the benefits and opportunities for public health research that arise from the linkage of these types of data.

1.2. WHY LINK SURVEY DATA WITH ADMINISTRATIVE DATA?

When linking survey data and administrative data, the advantages of the different data sources are combined while limitations of individual data sources can be compensated for. These synergy effects create an enriched body of data that forms the basis for answering new research questions (24). Indeed, health survey data are collected to monitor health status, well-being, health behaviour and other health determinants, and health care access, while administrative data are collected for other purposes than research, depending on the type of administrative data (e.g. the primary purpose of hospital discharge data is the financing of the hospitals; health insurance data are a tool to implement reimbursement of healthcare costs, but mortality data are collected for health monitoring).

Linking administrative data to population survey datasets provides important advantages:

1. It reduces respondents' burden by reducing the number of questions that need to be asked or by allowing some complex, detailed and uninteresting questions to be replaced by questions which respondents find more interesting or salient (4). Reducing respondent burden improves the quality of

the data collected by avoiding unanswered questions (or missing values in the dataset), for example.

2. It provides a means of enriching survey datasets with additional data not collected directly from the survey participants, offering vital information on their health outcomes. The enriched datasets provide opportunities for research that may not have otherwise been possible by allowing the exploration of new hypotheses not foreseeable using independent datasets (4,10,24–29).
3. It reduces the cost of obtaining additional information from the survey participants, given the expense of active follow-up procedures (29).
4. It offers a significant increase in the number of auxiliary variables that may be used to assess or adjust for non-response bias in survey data (4).
5. It serves to validate self-reported information.
6. It offers cost-effective means to maximize the use of existing publicly funded data collections.
7. It lays the groundwork for multidisciplinary health-research initiatives involving investigators from numerous fields, such as public health, epidemiology, pharmacology, economics and policy, owing to the combination of a wide range of information, such as health status, diagnosis, risk factors, use of health care and services, and socioeconomic background, at general population level.

1.3. PROOF OF CONCEPT IN LINKING HEALTH SURVEY AND ADMINISTRATIVE DATA

As previously mentioned, the linkage between survey data and administrative data has been used for a number of purposes. Here are a few examples.

1.3.1. Complementing survey data

Linkage with administrative data constitutes a powerful and cost-effective method for complementing survey data, specifically from cohort studies (30–33).

For example, in Germany, the lidA- leben in der Arbeit, a cohort study on work, age and health, utilizes a linkage between survey data and claims data from a large amount of statutory health insurance data (24).

In France, CONSTANCES is a very large “generalist” population-based cohort designed for health research and for providing public health information, which collects information from a representative sample of the French population, aged 18-69 years. The inclusion of participants takes place in a health examination centre of the insurance scheme, where they are given a medical examination and asked to complete a questionnaire about their health, their lifestyle and occupational history. A biobank (for blood and urine storage) is constituted. The follow-up includes an annual self-administered questionnaire, a medical examination every five years, and the linkage to various administrative data such as the SNIIRAM (national inter-scheme health insurance information system), the CNAV (National Old-Age Insurance Fund) and CépiDc (System for Automated Coding of Causes of Death) databases. Currently (May 2023), more than 200,000 participants are included in the cohort (31,32).

The Cohort of Norway (CONOR) uses the Norwegian unique identification numbers to link health survey data from consenting participants to administrative data (e.g. national health registries, drug prescription, disease registers, census), and thereby help build a nationally representative multipurpose cohort. This database has been used in several studies. For instance, Riise et al. (2021) assessed the association between casual blood glucose level and subsequent cardiovascular disease (CVD) and mortality among community-dwelling adults without a diagnosis of diabetes (34).

In Canada, the Canadian Longitudinal Study on Aging (CLSA) (35) and the Canadian Partnership for Tomorrow’s Health (CanPath) (36) are two nationwide longitudinal cohort studies that collect information on lifestyle and behaviours, health outcomes, social and physical environments. To complement information collected in each cohort and for passive follow-up of participants, both CLSA and CanPath plan to link consenting participants’ data to the information collected in provincial administrative health databases such as vital statistics, hospitalisations, physician billing, and drug prescription (35,36).

1.3.2. Building longitudinal studies

Survey data can be linked to one or more administrative data collections to form longitudinal population data that can be used for different research purposes. In 2020, Druschke et al. realised a data linkage of primary data obtained via a postal questionnaire to parents/caregivers of children born between 2007-2013 (aged 7 to 13 years - EcoCare-Pin birth cohort (37)) with two secondary data sources: 1) health insurance data (from 2007 to 2013), and 2) medical data from kindergarten- and school-entry examinations of Saxon health authorities. This linked longitudinal data collection enabled investigating the short- and long-term consequences of preterm birth with regard to parental stress, parent-child relationship, family and child quality of life, child development, and healthcare utilisation including costs (38).

The Integrated Data Infrastructure (IDI) database in New Zealand (NZ) consists of a central spine and many nodes (collections of datasets linked to the spine). The IDI spine is intended to capture the 'ever-resident' NZ population and is itself the result of a linkage between three key datasets: taxation from 1999 onwards, NZ births from 1920 onwards, and long-term visa approvals from 1997 onwards. Nodes are collections of datasets that share a common identifier and are usually collected by the same agency. For example, the people and communities node includes labour force and social surveys conducted by Statistics New Zealand, and the health node includes datasets such as pharmaceutical dispensing, lab tests, and hospital discharges. The linked data form a national-level longitudinal dataset that can be used for research, policy development and national statistical reporting (39–41).

The Western Australia Data Linkage System (WADLS), an international leader in data linkage, has now managed to link up to 40 years of data from over 30 population-based research surveys as well as administrative datasets (e.g. births, deaths, hospital inpatients, electoral rolls) covering the 2 million inhabitants of Western Australia (10). The WADLS has supported over 400 studies with over 250 journal publications and 35 graduate research degrees (10).

1.3.3. Validating survey information

Furthermore, linking survey data with administrative data is a well-established method for external validation of survey-based information (4,26,28,39,42–44), as is demonstrated in the following examples:

Hafferty et al. (2017) used survey-record data linkage to assess the validity of self-reported medication use against national prescribing data in Scotland and found a very good agreement for antidepressants and antihypertensives but moderate-poor agreement for mood stabilizers (45).

Hall et al. (2004) studied the validity of self-reported screening for prostate cancer and colorectal cancer in the United States using medical records as gold standard. The authors concluded that there was an overreporting of screening using self-reported data, making those data less appropriate to evaluate progress towards reaching national goals for prevention behaviours (42).

Richardson et al. (2013) assessed the agreement between interview-ascertained medication use and pharmacy records using the linkage between the Irish Longitudinal study on Ageing with pharmacy dispensing records. They concluded that ascertaining medication use via patient interview seems a valid method for most medication classes and also captures nonprescription and supplement use. However, topical medications and medications only used when needed may be underreported (43). Another study assessed the agreement between the results of a respiratory health survey conducted in Montreal on children aged 6 months to 12 years and the "*Regie de l'assurance maladie du Quebec*" (RAMQ, Quebec health insurance board) database in terms of the diagnosis of asthma and medical services use. The authors found moderate agreement between the two data sources for the diagnosis of asthma when a definition requiring 2 diagnoses in the RAMQ database was used (46).

1.3.4. Addressing methodological issues

Linked survey and administrative data have been used in studies for bias assessment. Indeed, Gorman et al. (2014) assessed the representativeness (selection bias) of population-sampled health surveys on alcohol-related outcomes through linkage to administrative data (47). Meyer et al. (2021) combined administrative and survey data

to improve income measurement (48). Morgan et al. (2020) also used a linkage between the national survey on adolescents' health and well-being in Wales with routine datasets (e.g. general practice, inpatient, and outpatient records) to assess the impact on overall parental consent rates on study completion and sample representativeness. The authors concluded that introducing data linkage consent within a national survey of adolescents had no impact on study completion rates. However, students consenting to data linkage, and those successfully linked, differed from non-consenting students on several key characteristics, raising questions concerning the representativeness of linked cohorts (49).

The linkage of survey data with administrative data can also serve for non-response analyses. For example, Linnenkamp et al. (2020) used the linkage between a cross-sectional study on depression among patients with diabetes and the German statutory health insurance data to evaluate whether non-response is a potential source of bias within a study. The authors found differences in age, sex, diabetes treatment and medication use between respondents and non-responders, which might bias the results, but did not find differences in terms of depression (50).

1.3.5. Addressing specific research questions

Linking survey data with administrative health data can be a useful population-based predictive tool or to study specific research questions. For example, Domhoff et al. (2023) will perform and evaluate the linkage of German Care Needs Assessment data with statutory health insurance claims data. The resulting dataset should enable the identification of factors in health care predicting the time between the onset of long-term care dependency and the admission to a nursing home in Germany in subsequent analyses (51). Using health survey data linked to administrative health services data, the Institute for Clinical and Evaluative Sciences (ICES) researchers in Ontario, Canada, developed and validated an algorithm for population-based prediction of diabetes - the Diabetes Population Risk Tool (DPoRT) that accurately predicts diabetes risk in a population (52).

Using the linkage of Canadian Community Health Survey (CCHS) with medical claim data, Rosella et al. (2014) investigated a wide range of individual-level characteristics that are associated with community-dwelling high-cost users. They found that high-

cost user status was strongly associated with being older, having multiple chronic conditions, and reporting poorer self-perceived health. The authors further found that high-cost users tended to be of lower socio-economic status, former daily smokers, physically inactive, current non-drinkers, and obese (53). Saunders et al. (2021) made use of a linkage of the 2015 Early Development Instrument (EDI) cycle and health administrative data to measure medical and social risk factors for early developmental vulnerability in Ontario, Canada. The authors highlighted the relative contribution of medical and social risk factors to developmental vulnerability and poor school achievement (54). Using the linkage of the Canadian Community Health Survey (CCHS) to hospital, physician and medication data, Lemstra et al. (2009) compared health care utilisation rates and costs between income groups in Saskatoon, Canada. They concluded that residents with lower income are responsible for disproportionate usage of hospitals, physicians and medications; due mainly (but not entirely) to higher disease prevalence (55).

1.3.6. In short, linking health survey and administrative data is an approach that benefits both data sources

From the above, it is clear that the linkage of survey data with administrative data is well established. For population health monitoring, one of the most important and tailored sources of information are population-sampled surveys, such as health interview surveys (HIS). A HIS plays an important role in shaping the development, implementation, and evaluation of public health policy and practice (47). By means of an HIS, health data can be collected that are ‘unique’ (no other way to collect such data) such as data related to mental health, self-perceived health or health behaviours, and it is possible to collect a range of information at the same time. However, an HIS can be expensive and time-consuming depending on the data collection approaches (e.g. face-to-face data collection is more expensive and time-consuming than online data collection). In addition, an HIS can be affected by different types of bias such as selection bias, recall bias or social desirability bias (24,56), which may impede valid inference. Some areas of the HIS are more prone to a specific type of bias than others. For example, events such as contacts with healthcare providers, preventive care or medication use are more likely to be biased by memory, while health habits such as smoking or alcohol consumption are more likely to be

biased by social desirability. While recall bias can be resolved by replacing HIS data with health administrative data, bias due to social desirability (e.g. smoking) cannot be resolved by linkage. Next to survey data, researchers are increasingly considering the secondary use of available health data, i.e. re-using data that were firstly gathered for a different purpose. As these data have already been collected, secondary data can theoretically be accessed easily in a quick and resource-efficient way. In addition to their cost efficiency in terms of data collection, secondary data offer additional advantages, depending upon their nature and source (38,57). For instance, administrative record data, such as data obtained from health insurance, from primary care files or hospital information systems, from disease-specific registers or from the mortality database (7,57), are increasingly used as secondary data for public health research purposes. Valid information on health and health care use is essential to accommodate health policies to the needs of the population. Health care information from records is usually considered as more accurate and reliable than self-reported information obtained in the context of an HIS. However, registered data are primally collected for administrative purposes and are not always suitable for epidemiological research. Moreover, they also have some limitations, as they may be incomplete and depend on the method by which they are collected (electronically or not). Finally, administrative data can be very complex, therefore requiring inside knowledge and clearly documented metadata to fully understand them.

1.4. CONTEXT OF THE LINKAGE BETWEEN THE BELGIAN HEALTH INTERVIEW SURVEY DATA AND THE BELGIAN COMPULSORY HEALTH INSURANCE DATA (HISLINK)

In this thesis we focus on the linkage of two major health data sources in Belgium: the Belgian Health Interview Survey (BHIS) and the Belgian Compulsory Health Insurance (BCHI).

The BHIS is a nationally representative, cross-sectional household interview survey that serves as an important source of information on the population's health and health-related behaviours in Belgium (58). The BHIS has been organised every 5 years since 1997 and aims to monitor the health status, well-being, health determinants and health care access of the population over time. One of the strengths of the BHIS is that it provides information on self-perceived health, lifestyle factors and behaviours, which can only be collected through surveys. It includes a multitude

of topics in the field of health, health determinants and care, but the increasing demand of the commissioners and stakeholders to inflate the content of the questionnaire faces a significant challenge. BHIS data are also used by external researchers in different fields, who request more and more complete and detailed data to perform more in-depth analyses. However, there is of course a limit to the length of the survey questionnaire, as more tedious interviews lead to an increasing burden for both the interviewers and the interviewees. The main impact of this phenomenon is a loss of the quality of the collected data due to fatigue and the reduction in confidence in the conclusions drawn from the survey (29) due to increase of the non-response rate. There is therefore a pressing need to decrease the burden of the survey for the interviewers and the interviewees in order to collect good-quality data. One approach to reduce the length of the BHIS questionnaire is to replace or substitute, where possible, information traditionally obtained from the BHIS with information existing in other data sources, such as administrative records. This is particularly the case for information on medical consumption.

The BCHI also has an important role in collecting health-related data. Indeed, In Belgium, there is compulsory health insurance, which is a source of exhaustive and detailed data on the reimbursed health expenses of over 98% of the total population. Almost every citizen is a member of one of the seven "*ziekenfonds*" or "healthcare or sickness funds" (compulsory-health insurance organisations). Since 2002, the InterMutualistic Agency (IMA), an overarching national organisation, collects and manages data on all Belgian citizens from these healthcare funds (hereinafter referred to as BCHI data). Therefore, the BCHI is the most important administrative data source regarding population healthcare consumption in Belgium. These data are widely exploited by important actors in the health field, such as the National Institute for Health and Disability Insurance (NIHDI), the Belgian Health Care Knowledge Centre (KCE), the Belgian Federal Planning Bureau and the healthcare insurers, for reimbursement purposes, assessment and planning of healthcare costs. In addition, BCHI data have also been used for specific studies, beyond their initial intended use (secondary use). Furthermore, since BCHI data registrations are usually standardised and continuously collected, they enable trend analyses and longitudinal studies (59). However, the BCHI data also have some shortcomings. Because they are collected for administrative purposes, they might not include all the relevant information to

answer specific research questions. For instance, sociodemographic, clinical information and mental health information are limited or non-existent.

As already discussed, both the BHIS and BCHI data have their strengths and shortcomings. While BHIS data are self-reported and subject to bias, BCHI data tend to be more valid but lack information on important health determinants, such as socio-demographic and lifestyle information. By linking survey data with administrative data, more comprehensive and high-quality data can be included in the BHIS database without increasing the workload for the interviewers and the interviewees. The BHIS data have been previously linked with other data sources such as mortality data (60) and census data (60,61).

In 2012, the NIHDI commissioned a feasibility study on a possible linkage between BHIS 2008 and BCHI data from 2007 (or 2005 for some specific cases) to 2010 (further referred to as HISlink). The main objective of this pilot study was to serve three specific research purposes: (a) to explore regional differences in healthcare consumption in more depth; (b) to assess the validity of healthcare-consumption-based chronic disease indicators against self-reports of chronic diseases; (c) to estimate the cost to Belgian health insurance if some groups of non-reimbursed medicines (analgesics, laxatives and calcium supplements) were to be reimbursed (62). Moreover, Sciensano took advantage of the linkage opportunity to request two additional objectives: (a) to substitute where possible BHIS-based data with BCHI data (in order to decrease the BHIS burden and to have less biased data) and (b) to enrich the BHIS dataset. Once these two datasets were linked, two more studies were carried out on the linked data (56,63). The success of the pilot study allowed to foresee a systematic linkage of the data sources, the “HISlink project” starting in 2017, i.e. between each wave of BHIS and corresponding BCHI data. The project is specifically meant to respond to policy-relevant questions raised by NIDHI, who is the commissioner. The HISlink project takes advantage of the strengths of both data sources to be used synergistically and provides opportunities for new and advanced research. While the BHIS data on medical consumption could be subject to recall bias and thus be inaccurate, so prone to substitution by BCHI data, it is a reliable source for detailed information on sociodemographic data, health-related behaviour and mental health. On the other hand, the BCHI data gathers elements that cannot be collected by means of a survey (e.g. costs of health care). In this perspective, linking

both data sources yields a 'richer' database. However, such a linkage has its own challenges and considerations that need to be taken into account. The challenges may vary according to the context such as the need of linkage consent and the applicable data protection requirements. Within the framework of HISlink, data from two BHIS waves have been linked to BCHI data: BHIS 2013 and BHIS 2018. The linkage procedure with data from BHIS 2023 is under preparation.

1.5. PRINCIPAL OBJECTIVES AND RESEARCH QUESTIONS

Based on the use case of the HISlink, the overarching aim of this thesis is to investigate the potential benefits and opportunities of linking survey data with health insurance data for public health research.

The following parts and questions will be addressed:

1. To explore the fundamental concepts of data linkage, a literature review was undertaken to cover the following questions: What is data linkage? What are commonly the types of linked data? What methods have been used to link data? What are the challenges and the legal issues? How to assess the quality of linked data?

Then, the following two research questions were examined:

2. Despite the increasing availability of health and health-related administrative data, self-reported information obtained through questionnaires in national health interview surveys (HISs) such as BHIS remains an important source of health information. An HIS provides information on a wide range of health-related topics, measured at the same time and at the level of the total population (including people not using healthcare services) from the perspective of the individual. HISs are extensively used to make comparisons between population groups and between countries, and to assess time trends. In the European Union, all Member States (MS) collect data on the health status, provision of healthcare, health determinants and socioeconomic situation to feed a common European Health Interview Survey (EHIS) (64).

   EHIS is an example of use of survey data for comparison between countries and across time periods. Harmonised data collected through EHIS serve to

construct European health indicators that are of key importance to the national- and European-level health policies and play an important role in comparisons between MS. The data collected through the HISs are often used for reporting to international instances such as WHO, UN and OECD where they are used to feed important reports such as the OECD Health at a Glance report (65), the Health System Performance Assessment (HSPA) (66) and the European Health Report published jointly by the WHO's Regional Office for Europe and the European Commission (64). Furthermore, national HISs-based information is widely available and used by many people (policymakers, researchers, healthcare professionals, patient organisations, journalists, etc.) and for many purposes (67). It is therefore important to ensure that HIS-based information is valid. However, despite numerous studies on the validity of HIS-based data (43,45,46,67,68), the validity of self-reported information remains a cause for concern. Therefore, research on the validity of HIS-based estimates is relevant and needs to be continuously updated. Data linkage can play a crucial role in obtaining further insights about the validity of self-reported information. The first research question is therefore: ***To what extent can linked data be used to assess data validity?***

3. Researchers are increasingly faced with a greater demand for data in different areas of public health and research questions that require more comprehensive data, or even data from multiple sources. Previous studies have highlighted the need to implement linkage between HISs and administrative data sources (67,69) as a part of the solution to obtain a complete picture of population health and health-related information without increasing survey workloads.

   The BHIS and BCHI data are complementary and represent valuable sources of information in the Belgian health information system, but with their own strengths and limitations. Linking BHIS and BCHI data will result in a richer database that offers new research opportunities for public health authorities. Therefore, an important aim of this thesis is to present use cases that demonstrate the added value of linked survey and administrative data to explore policy-relevant research questions that cannot be sufficiently

investigated by each of the databases separately. The second research question is therefore the following: ***To what extent can linked data be used to respond to policy-relevant questions which cannot be addressed with each of the sources separately?***

Although the objectives of this thesis are quite broad, offering a wide range of research possibilities, only a limited number of topics were selected to meet them. This selection was guided by the relevance of the topics for public health (relevance for the commissioner of the linkage, relevance with respect to societal challenges), and the feasibility in relation to the information available in both databases.

Hence, the first research question was explored for three topics: self-reported mammography uptake, chronic diseases, and polypharmacy; and the second research question was explored for two policy-relevant research questions: What are predictors of nursing-home admission among the older population? What is the mediating effect of health literacy in the relationship between socioeconomic status and health outcomes.

1.6. OUTLINE OF THE THESIS

The outline of this thesis is summarised below.

Chapter 2 related to the background of data linkage, outlines some fundamental concepts relating to data linkage. It provides the definitions of data linkage, a broader description of common types of databases involved in linkages and presents an overview of data linkage methods. In addition, it describes the challenges, privacy concerns and legal issues related to data linkage as well as practical considerations to be taken into account when planning to link databases. Finally, the chapter focuses on quality assessment of linked data and linked data validation. However, in this chapter, no attempt will be made to provide a complete review of the literature on data linkage. Therefore, it is beyond the scope to comprehensively review all the data linkage aspects. Only a broader overview of the related aspects mentioned above will be provided. The interested reader may refer to Dusetzina et al. (2014) (70) for an expanded review of the literature. In addition, a large collection of work on record linkage by various authors with extensive references is presented (71–74). Chapter 2 is based on a literature review and is useful for understanding the next chapter 3 on

the HISlink implementation in terms of the linkage methods, the challenges and privacy issues encountered, the quality assessment and the validation of the linked data.

Chapter 3 provides a concise description of data sources, including the BHIS data, the BCHI data and the random sample of the BCHI and focuses on the HISlink as a use case of linking survey data with health insurance data. It presents the practical implementation of the HISlink, the main challenges and privacy issues encountered, the main outcomes in terms of linked data, and useful recommendations for future data linkages.

On the basis of three cases, chapter 4 shows how linked data can be used in validation studies in the presence of gold standard, in comparison of data sources when there is not a gold standard and demonstrates the added value of using different but complementary data sources to study the same research question with the same study population. Indeed, the linked data offers opportunities to answer methodological questions on the validity of survey information, such as the validity of self-reporting information. For instance, data on the mammography uptake is usually based on self-reports in population-based surveys such as BHIS. However, the validity of self-reported information through surveys is a concern, due to the associated potential reporting bias. To gain further insights into the validity of self-reported mammography uptake in Belgium, in the first paper related to this chapter, we assessed the selection and reporting biases of BHIS-based estimates in the target group (women aged 50–69 years) using reimbursement data for mammograms taken from the BCHI.

Currently, the estimation of the prevalence of many chronic diseases (CDs) in Belgium is still often based on self-reported BHIS data. On the NIHDI's initiative, we evaluated whether BCHI data can be used to ascertain the prevalence of CDs in the Belgian population. For this purpose, in the second paper of this chapter, the linkage was utilised to study the agreement between BHIS-based diagnosis and pseudo-diagnosis based on health consumption for a selected number of CDs.

The third paper in this chapter demonstrates the use of linkage to show how polypharmacy can be addressed from different angles and how this yields complementary information. More specifically, this paper explores the agreement

between polypharmacy (use or prescription of ≥5 medicines) and excessive polypharmacy (≥10 medicines) between both sources in the older general population in Belgium and assesses the relative merits of each data source.

Other BHIS information as a good candidate for validation as compared to BCHI data are information on contact with healthcare providers. However, the validity of these indicators has already been tested in the framework of a previous study (67) and is therefore not discussed in the frame of this thesis.

The studies presented in chapter 5 and chapter 6 both show the potential of linked data to study policy-relevant research questions which cannot or can only be investigated less precisely with one database only. Although both studies address the added value of linked data in terms of policy-research questions, they use different aspects of linked data that deserve to be separated (longitudinal and cross-sectional aspects).

Chapter 5 concerns the use of linked survey data with administrative data for longitudinal study. The linked data not only increase the number of variables, but also make it possible to track the healthcare consumption of BHIS participants over time. Tracking BHIS participants up to 5 years after the survey, research questions can be addressed that require a longitudinal design. The paper in this chapter estimates the cumulative risk of nursing-home admission (NHA) among the older population of 65+ years at 1 year, 3 years and 5 years of follow-up and its predictors in Belgium.

Chapter 6 concerns a use case showing the added value of linked data in answering policy-relevant questions. The NIHDI is interested in exploring the extent to which health literacy (HL) can mediate the relationship with SES, as measured by education, household income and health related outcomes in areas that are of high interest to policymakers, such as health prevention, health behaviour, health status including mental health. From a policy perspective, estimating the total causal effect that is due to the mediation of HL could help to set up interventions to reduce socioeconomic health disparities. The related paper to this chapter explores the mediating effects of HL on the relationship between education, income and a selected health related outcomes in varying domains: 1) health behaviour (physical activity, diet, alcohol and tobacco consumption), 2) health status (perceived health status, mental health), 3)

use of medicines (purchase of antibiotics), and 4) use of preventive care (preventive dental care, influenza vaccination, breast cancer screening).

Chapter 7 is based on a summary paper that provides an overview of the methodology used in the HISlink, the principal challenges and privacy issues encountered and the main outcomes in terms of linked data, and useful recommendations for future data linkages.

Finally, chapter 8 briefly summarises the research problem and the main findings of the thesis. It then reviews the strengths and limitations, followed by future perspectives, implications and recommendations, and ends with a final conclusion.

Table 1.1 provides an overview of the thesis objectives, chapters and the related publications.

Table 1.1: Overview of the thesis structure

Thesis parts and objectives	**Thesis chapters**	**Papers and related chapter**
Background and descriptive information	Chapter 1: General introduction	
	Chapter 2: Introducing data linkage	
	Chapter 3: Data sources and implementation of HISlink	Paper 6 (Summary paper): Linking health survey data with health insurance data: methodology, challenges, opportunities and recommendations for public health research. An experience from the HISlink project in Belgium
	Chapter 7: HISlink: methodology, challenges, opportunities and recommendations for public health research	
Objective 1: Validation (and comparison)	Chapter 4.1: validation	Paper 1: Validity of self-reported mammography uptake in the Belgian health interview survey: selection and reporting bias
	Chapter 4.2: comparison	Paper 2: Comparing administrative and survey data for ascertaining chronic disease prevalence

	Chapter 4.3: comparison	Paper 3: Assessing polypharmacy in the older population: comparison of a self-reported and prescription-based method
Objective 2. Added value for policy-relevant questions	Chapter 5: added value (longitudinal study)	Paper 4: Predictors of nursing home admission in the older population in Belgium: a longitudinal follow-up of health interview survey participants
	Chapter 6: added value (additional policy-relevant questions)	Paper 5: Does health literacy mediate the relationship between socioeconomic status and health related outcomes in the Belgian adult population?
General discussion and recommendations	Chapter 8: General discussion and recommendations	

1.7. CONCLUSIONS

Linking survey data with health administrative data is increasingly used in public health research. Such linkage offers new opportunities for research into the use of health services and public health. Building on the experience of the linkage between BHIS and BCHI data sources, this thesis will summarise useful background information on data linkage and the practical implementation of linking data. The use of linked data to validate self-reported information or to compare complementary data sources will be demonstrated. Additionally, the added value of the linked data in terms of answering policy-relevant questions will be addressed. Finally, some recommendations for future linkages will be formulated.

1.8. BIBLIOGRAPHY

1. Centre for Health Record Linkage (CHeReL). New South Wales (NSW) Government Website - Centre for Health Record Linkage. [cited 2023 Feb 9]. How record linkage works. Available from: https://www.cherel.org.au/how-record-linkage-works#:~:text=How%20record%20linkage%20works,of%20health%20events%20for%20individuals.
2. Brook EL, Rosman DL, Holman CDJ. Public good through data linkage: measuring research outputs from the Western Australian Data Linkage System. Australian and New Zealand Journal of Public Health. 2008 Feb;32(1):19–23.
3. Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang LC, Smith P, et al. GUILD: GUidance for Information about Linking Data sets†. Journal of Public Health. 2018 Mar 1;40(1):191–8.
4. Calderwood L, Lessof C. Enhancing Longitudinal Surveys by Linking to Administrative Data. In: Lynn P, editor. Methodology of Longitudinal Surveys [Internet]. Chichester, UK: John Wiley & Sons, Ltd; 2009 [cited 2020 Sep 29]. p. 55–72. Available from: http://doi.wiley.com/10.1002/9780470743874.ch4
5. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. Big Data & Society. 2017 Dec;4(2):205395171774567.
6. Holman CDJ, Bass AJ, Rouse IL, Hobbs MST. Population-based linkage of health records in Western Australia: development of a health services research-linked database. Australian and New Zealand Journal of Public Health. 1999 Oct;23(5):453–9.
7. Haneef R, Delnord M, Vernay M, Bauchet E, Gaidelyte R, Van Oyen H, et al. Innovative use of data sources: a cross-sectional study of data linkage and artificial intelligence practices across European countries. Arch Public Health. 2020 Dec;78(1):55.
8. Aldridge RW, Shaji K, Hayward AC, Abubakar I. Accuracy of Probabilistic Linkage Using the Enhanced Matching System for Public Health and Epidemiological Studies. Pacheco AG, editor. PLoS ONE. 2015 Aug 24;10(8):e0136179.
9. Green E, Ritchie F, Mytton J, Webber DJ, Deave T, Montgomery A, et al. Enabling data linkage to maximise the value of public health research data: Summary report. 2015
10. Holman CDJ, Bass JA, Rosman DL, Smith MB, Semmens JB, Glasson EJ, et al. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. Aust Health Review. 2008;32(4):766.
11. Antoni M. Linking survey data with administrative employment data: The case of the German ALWA survey. FDZ Methodenreport. 2011;12:2012.

12. Tew M, Dalziel KM, Petrie DJ, Clarke PM. Growth of linked hospital data use in Australia: a systematic review. Aust Health Review. 2017;41(4):394.
13. Young A, Flack F. Recent trends in the use of linked data in Australia. Aust Health Review. 2018;42(5):584.
14. Maret-Ouda J, Tao W, Wahlin K, Lagergren J. Nordic registry-based cohort studies: Possibilities and pitfalls when combining Nordic registry data. Scand J Public Health. 2017 Jul;45(17_suppl):14–9.
15. Ali MS, Ichihara MY, Lopes LC, Barbosa GCG, Pita R, Carreiro RP, et al. Administrative Data Linkage in Brazil: Potentials for Health Technology Assessment. Front Pharmacol. 2019 Sep 23;10:984.
16. Carrière G, Sanmartin C, Murison P. Using data linkage to report surgical treatment of breast cancer in Canada. Health Rep. 2018;29:3–8.
17. Harron K, Gilbert R, Cromwell D, van der Meulen J. Linking Data for Mothers and Babies in De-Identified Electronic Health Data. Gebhardt S, editor. PLoS ONE. 2016 Oct 20;11(10):e0164667.
18. Schmidt M, Schmidt SAJ, Adelborg K, Sundbøll J, Laugesen K, Ehrenstein V, et al. The Danish health care system and epidemiological research: from health care contacts to database records. CLEP. 2019 Jul;Volume 11:563–91.
19. Di Rico R, Nambiar D, Gabbe B, Stoové M, Dietze P. Patient-specific record linkage between emergency department and hospital admission data for a cohort of people who inject drugs: methodological considerations for frequent presenters. BMC Med Res Methodol. 2020 Dec;20(1):283.
20. Marriott JJ, Chen H, Fransoo R, Marrie RA. Validation of an algorithm to detect severe MS relapses in administrative health databases. Multiple Sclerosis and Related Disorders. 2018 Jan;19:134–9.
21. Lane T, Berecki-Gisolf J, Iles R, Collie A, Smith P. impact of workers' compensation benefit cessation on welfare and health service use: Protocol for a longitudinal controlled data linkage study. IJPDS [Internet]. 2021 May 12 [cited 2023 May 24];6(1). Available from: https://ijpds.org/article/view/1419
22. Huang N, Shih SF, Chang HY, Chou YJ. Record linkage research and informed consent: who consents? BMC Health Serv Res. 2007 Dec;7(1):18.
23. Chambers R, Banati P, McMaster NC. Opportunities and Challenges of Data Linkage for Longitudinal Surveys [Internet]. Workshop on The Future of the HILDA Survey - Opportunities and Challenges; 2017 Sep 7 [cited 2023 May 24]; Melbourne. Available from: https://www.unicef-irc.org/files/upload/documents/HILDA%20Linkage%20Presentation.pdf
24. March S. Individual Data Linkage of Survey Data with Claims Data in Germany—An Overview Based on a Cohort Study. IJERPH. 2017 Dec 9;14(12):1543.
25. Swart E, Stallmann C, Powietzka J, March S. Datenlinkage von Primär- und Sekundärdaten: Ein Zugewinn auch für die kleinräumige Versorgungsforschung in Deutschland? Bundesgesundheitsbl. 2014 Feb;57(2):180–7.

26. Hamood R, Hamood H, Merhasin I, Keinan-Boker L. A feasibility study to assess the validity of administrative data sources and self-reported information of breast cancer survivors. Isr J Health Policy Res. 2016 Dec;5(1):50.
27. Linkage Between Cohorts and Health Care Utilization Data: Meeting of Canadian Stakeholders workshop participants, Doiron D, Raina P, Fortier I. Linking Canadian Population Health Data: Maximizing the Potential of Cohort and Administrative Data. Can J Public Health. 2013 May;104(3):e258–61.
28. Zuvekas SH, Olin GL. Validating Household Reports of Health Care Use in the Medical Expenditure Panel Survey. Health Services Research. 2009 Oct;44(5p1):1679–700.
29. Sakshaug JW, Couper MP, Ofstedal MB, Weir DR. Linking Survey and Administrative Records: Mechanisms of Consent. Sociological Methods & Research. 2012 Nov;41(4):535–69.
30. Gao L, Leung MTY, Li X, Chui CSL, Wong RSM, Au Yeung SL, et al. Linking cohort-based data with electronic health records: a proof-of-concept methodological study in Hong Kong. BMJ Open. 2021 Jun;11(6):e045868.
31. Constances [Internet]. [cited 2023 May 25]. Available from: https://www.constances.fr/index_EN.php
32. CONSTANCES team, Zins M, Goldberg M. The French CONSTANCES population-based cohort: design, inclusion and follow-up. Eur J Epidemiol. 2015 Dec;30(12):1317–28.
33. Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. Annals of Human Biology. 2020 Feb 17;47(2):218–26.
34. Riise HKR, Igland J, Sulo G, Graue M, Haltbakk J, Tell GS, et al. Casual blood glucose and subsequent cardiovascular disease and all-cause mortality among 159 731 participants in Cohort of Norway (CONOR). BMJ Open Diab Res Care. 2021 Feb;9(1):e001928.
35. CLSA and HDRN Canada partner to enable data linkage [Internet]. [cited 2023 May 25]. Canadian Longitudinal Study on Aging (CLSA). Available from: CLSA and HDRN Canada partner to enable data linkage | Canadian Longitudinal Study on Aging (clsa-elcv.ca)
36. CanPath, the Canadian Partnership for Tomorrow's Health [Internet]. [cited 2023 May 25]. Available from: https://canpath.ca/
37. Rüdiger M, Heinrich L, Arnold K, Druschke D, Reichert J, Schmitt J. Impact of birthweight on health-care utilization during early childhood – a birth cohort study. BMC Pediatr. 2019 Dec;19(1):69.
38. Druschke D, Arnold K, Heinrich L, Reichert J, Rüdiger M, Schmitt J. Individual-Level Linkage of Primary and Secondary Data from Three Sources for Comprehensive Analyses of Low Birthweight Effects. Gesundheitswesen. 2020 Mar;82(S 02):S108–16.

39. Kvalsvig A, Gibb S, Teng A. Linkage error and linkage bias: A guide for IDI users. University of Otago. 2019.
40. Atkinson J, Blakely T. New Zealand's Integrated Data Infrastructure (IDI): Value to date and future opportunities: IJPDS (2017) Issue 1, Vol 1:105, Proceedings of the IPDLN Conference (August 2016). IJPDS [Internet]. 2017 Apr 18 [cited 2023 May 25];1(1). Available from: https://ijpds.org/article/view/124
41. Jones C, McDowell A, Galvin V, Adams D. Building on Aotearoa New Zealand's integrated data infrastructure. Harvard Data Science Review. 2022;4(2).
42. Hall HI, Van Den Eeden SK, Tolsma DD, Rardin K, Thompson T, Hughes Sinclair A, et al. Testing for prostate and colorectal cancer: comparison of self-report and medical record audit. Preventive Medicine. 2004 Jul;39(1):27–35.
43. Richardson K, Kenny RA, Peklar J, Bennett K. Agreement between patient interview data on prescription medication use and pharmacy records in those aged older than 50 years varied by therapeutic group and reporting of indicated health conditions. Journal of Clinical Epidemiology. 2013 Nov;66(11):1308–16.
44. Li J, Cone JE, Alt AK, Wu DR, Liff JM, Farfel MR, et al. Performance of Self-Report to Establish Cancer Diagnoses in Disaster Responders and Survivors, World Trade Center Health Registry, New York, 2001–2007. Public Health Rep. 2016 May;131(3):420–9.
45. Hafferty JD, Campbell AI, Navrady LB, Adams MJ, MacIntyre D, Lawrie SM, et al. Self-reported medication use validated through record linkage to national prescribing data. Journal of Clinical Epidemiology. 2018 Feb;94:132–42.
46. Plante C, Goudreau S, Jacques L, Tessier F. Agreement between survey data and Régie de l'assurance maladie du Québec (RAMQ) data with respect to the diagnosis of asthma and medical services use for asthma in children. Chronic Dis Inj Can. 2014 Nov;34(4):256–62.
47. Gorman E, Leyland AH, McCartney G, White IR, Katikireddi SV, Rutherford L, et al. Assessing the Representativeness of Population-Sampled Health Surveys Through Linkage to Administrative Data on Alcohol-Related Outcomes. American Journal of Epidemiology. 2014 Nov 1;180(9):941–8.
48. Meyer BD, Mittag N. Combining administrative and survey data to improve income measurement. Administrative Records for Survey Methodology. 2021;297–322.
49. Morgan K, Page N, Brown R, Long S, Hewitt G, Del Pozo-Banos M, et al. Sources of potential bias when combining routine data linkage and a national survey of secondary school-aged children: a record linkage study. BMC Med Res Methodol. 2020 Dec;20(1):178.
50. Linnenkamp U, Gontscharuk V, Brüne M, Chernyak N, Kvitkina T, Arend W, et al. Using statutory health insurance data to evaluate non-response in a cross-sectional study on depression among patients with diabetes in Germany. International Journal of Epidemiology. 2020 Apr 1;49(2):629–37.

51. Domhoff D, Seibert K, Stiefler S, Wolf-Ostermann K, Peschke D. Data linkage of German statutory health insurance claims data and care needs assessments preceding a population-based cohort study on nursing home admission. BMJ Open. 2022 Jun;12(6):e063475.
52. Rosella LC, Manuel DG, Burchill C, Stukel TA, for the PHIAT-DM team. A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT). Journal of Epidemiology & Community Health. 2011 Jul 1;65(7):613–20.
53. Rosella LC, Fitzpatrick T, Wodchis WP, Calzavara A, Manson H, Goel V. High-cost health care users in Ontario, Canada: demographic, socio-economic, and health status characteristics. BMC Health Serv Res. 2014 Dec;14(1):532.
54. Saunders NR, Janus M, Porter J, Lu H, Gaskin A, Kalappa G, et al. Use of administrative record linkage to measure medical and social risk factors for early developmental vulnerability in Ontario, Canada. IJPDS [Internet]. 2021 Feb 11 [cited 2022 Mar 3];6(1). Available from: https://ijpds.org/article/view/1407
55. Lemstra M, Mackenbach J, Neudorf C, Nannapaneni U. High health care utilization and costs associated with lower socio-economic status: results from a linked dataset. Canadian Journal of Public Health. 2009;100:180–3.
56. Van der Heyden J, Charafeddine R, De Bacquer D, Tafforeau J, Van Herck K. Regional differences in the validity of self-reported use of health care in Belgium: selection versus reporting bias. BMC Med Res Methodol. 2016 Dec;16(1):98.
57. March S, Andrich S, Drepper J, Horenkamp-Sonntag D, Icks A, Ihle P, et al. Good Practice Data Linkage (GPD): A Translation of the German Version. IJERPH. 2020 Oct 27;17(21):7852.
58. Sciensano. HIS - Health Interview Survey [Internet]. [cited 2023 May 25]. Available from: https://www.sciensano.be/en/projects/health-interview-survey
59. Maetens A, De Schreye R, Faes K, Houttekier D, Deliens L, Gielen B, et al. Using linked administrative and disease-specific databases to study end-of-life care on a population level. BMC Palliat Care. 2016 Dec;15(1):86.
60. Charafeddine R, Berger N, Demarest S, Van Oyen H. Using mortality follow-up of surveys to estimate social inequalities in healthy life years. Popul Health Metrics. 2014 Dec;12(1):13.
61. Van der Heyden J, De Bacquer D, Tafforeau J, Van Herck K. Reliability and validity of a global question on self-reported chronic morbidity. J Public Health. 2014 Aug;22(4):371–80.
62. Mimilidis Hélène, Demarest Stefaan, Tafforeau Jean, Van der Heyden Johan. Projet de couplage de données issues de l'Enquête de Santé 2008 et des Organismes Assureurs. Bruxelles, Belgique; 2014 Mai. Report No.: D/2014/2505/32.

63. Van der Heyden J, Van Oyen H, Berger N, De Bacquer D, Van Herck K. Activity limitations predict health care expenditures in the general population in Belgium. BMC Public Health. 2015 Dec;15(1):267.
64. Hintzpeter B, Finger JD, Allen J, Kuhnert R, Seeling S, Thelen J, et al. European health interview survey (EHIS) 2–background and study methodology. Journal of Health Monitoring. 2019;4(4):66.
65. OECD, European Union. Health at a Glance: Europe 2022: State of Health in the EU Cycle [Internet]. OECD; 2022 [cited 2023 Aug 10]. (Health at a Glance: Europe). Available from: https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-europe-2022_507433b0-en
66. Perić N, Hofmarcher-Holzhacker MM, Simon J. Health system performance assessment landscape at the EU level: a structured synthesis of actors and actions. Archives of Public Health. 2017;75:1–10.
67. Van der Heyden J. Validity of the Assessment of Population Health and Use of Health Care in a National Health Interview Survey [Internet]. [Ghent, Belgium]: Ghent University - Faculty of medicine and health sciences; 2017 [cited 2023 Feb 9]. Available from: https://biblio.ugent.be/publication/8523878
68. Short ME, Goetzel RZ, Pei X, Tabrizi MJ, Ozminkowski RJ, Gibson TB, et al. How Accurate are Self-Reports? Analysis of Self-Reported Health Care Utilization and Absence When Compared With Administrative Data. Journal of Occupational & Environmental Medicine. 2009 Jul;51(7):786–96.
69. Braekman E. Going online with the health interview survey? Assessing the effect of data collection mode on participation, measurements and costs in a Belgian context [Internet]. Antwerp; 2020 [cited 2023 May 29]. Available from: https://repository.uantwerpen.be/docman/irua/703fcc/172527.pdf
70. Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR. Linking data for health services research: a framework and instructional guide. 2014.
71. Boyd JH, Ferrante AM, O'Keefe CM, Bass AJ, Randall SM, Semmens JB. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. BMC Health Serv Res. 2012 Dec;12(1):480.
72. Ranbaduge T. A Scalable Blocking Framework for Multidatabase Privacy-preserving Record Linkage. The Australian National University; 2018.
73. Harron K, Mackay E, Elliot M. An introduction to data linkage. 2016;
74. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. International Journal of Epidemiology. 2017 Oct 1;46(5):1699–710.

CHAPTER 2. INTRODUCING DATA LINKAGE

2.1. INTRODUCTION

The term record linkage was first used in 1946 when Dunn described linkage of vital records from the same individual (birth and death registrations) and referred to the process as “assembling the book of life”: «Each person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this Book, into a volume» (1). In fact, the ‘book of life’ for every individual in the world as described by Dunn contains pages covering the principal events of a life, such as the individual’s contacts with the health and social security systems from birth to death. This description of a volume containing a chronological history of significant life events from every aspect of a person’s lifetime provides a perfect picture of what record linkage can achieve, with each book containing a different story (2).

Data linkage is a method that brings together information that relates to the same individual, family, place or event from different data sources (3,4). Data linkage is also referred to as record linkage in computer science (5,6). Other synonyms such as ‘record matching’, ‘entity resolution’, ‘merge-purge’ (6), ‘data matching’, ‘entity resolution’, ‘co-reference resolution’ or ‘deduplication’ (7) are also used. The term data linkage will be used in this thesis.

Data linkage has become an increasingly used method for service evaluation and research (6). The increasing power of computers since the 1980s (8) as well as the development of computerised record linkage (9) played a crucial role in this progress. This digital era makes it possible to link even large data sets, such as large population-based linkage (8). Population-based linkage systems have been established in several countries around the world, including Australia, Canada, the UK and the Nordic countries (6,10–15). Data from different sources can be linked together and, depending on the contents of the databases, the best methods of linkage should be selected. Moreover, because data linkage has been increasingly used for research, there has been a growing interest in methodological issues associated with the creation and analysis of linked datasets (6,14,16–18). In the process of data linkage, researchers need to take into account several considerations. Bradley et al. (2010) identified five basic steps for linking databases:

(1) Identify the data sources that can be linked to answer a specific research question.

(2) Obtain the necessary approvals, including institutional ethics boards, regulatory authorities, and funding sources.

(3) Select the variables that will be used to link the databases and individually clean the datasets.

(4) Determine the best method for linking databases and develop algorithms accordingly.

(5) Evaluate the quality of the link between data sources (19).

A detailed description of the above five steps is beyond the scope of this thesis. Only the common types of data sources involved in linkages, data linkage methods including linkage variables, quality assessment of linked data and validation of linked data will be discussed. The necessary institutional approvals will not be discussed in this chapter as they vary considerably from country to country and, depending on the type of data to be linked and the organisations involved in the linkage (public or private organisations) and their interests, the available regulatory authorities may be bound by different laws. However, in the following chapter, the institutional permissions for the HISlink use case will be described. Next to those steps described by Bradley et al. (2010), this chapter also summarises the types of data linkage, describes the challenges, privacy concerns and legal issues related to data linkage as well as practical considerations to be taken into account when planning to link databases.

Because this thesis does not focus on data linkage per se, no attempt will be made to provide a complete review of the literature on data linkage. Therefore, it is beyond the scope to comprehensively review all the data linkage aspects. Only a broader overview of the related aspects mentioned above will be provided. The interested reader may refer to (20). for an expanded review of the literature. In addition, a large collection of work on record linkage by various authors with extensive references is presented (6,14,21–24).

The current chapter is based on literature review.

2.2. COMMONLY LINKED DATABASES WITHIN THE CONTEXT OF HEALTH RESEARCH

With the evolution of data linkage techniques, researchers are now able to link multiple and varying data sources within the context of health research – including surveys data and administrative data. Surveys data, whether cross-sectional or longitudinal, are usually collected for specific research purposes. Data from health interview surveys, health examination surveys and social surveys are the most common survey data that are linked with administrative data.

Administrative data are essentially collected for purposes other than research. They are usually collected for the purpose or in the process of service delivery, such as the provision of health care (e.g. hospital discharge data), to respond to legal requirements of registering particular events (e.g. births and deaths registration data) or to provide a particular service (21). Such data can be governmental or institutional. As with survey data, administrative data may be either longitudinal or cross-sectional in nature. Many administrative datasets store information by spell (i.e. by period), e.g. period of welfare benefit receipt or time spent in hospital. Such datasets are inherently longitudinal as successive spells for a given individual can be linked with each other so that change can be observed over time (25). Administrative data could be from multiple sources, either population-based or institutional-based and typically include healthcare administrative data, vital registrations systems data, census data, labour market and social protection data, financial data, environmental data, etc.

Healthcare administrative data are mainly collected for financial or clinical management purposes and are generated at every encounter with the health care system, e.g. in relation to a visit to a physician's office, a diagnostic procedure, an admission to hospital, or the reception of a prescription at a community pharmacy. The terms "healthcare utilisation data", "administrative healthcare billing records", "administrative claims data", or simply "claims data" are synonyms of "healthcare administrative data" (26). Common sources of health-related administrative data are health insurance claims data, hospital discharge data, prescription drugs data, medical records, disease-specific registries, etc.

Table 2.1 summarises the most common databases that are often linked and the key items they contain besides the individual's identification number (IDs).

Table 2.1: Commonly linked databases within the context of health research

Category	Description and relevant items	Examples	Website / Reference
(Health) Survey data	Contains self-reported information on health status, health care use, sociodemographic information, health behaviour and clinical measurements (blood pressure, height, weight, etc.) in case of examination surveys, etc.	Belgian Health Interview Survey (BHIS)	Health Interview Survey | sciensano.be
		National Health Interview Survey (NHIS), USA	NHIS - National Health Interview Survey (cdc.gov)
		National Health and Nutrition Examination Survey (NHANES), USA	NHANES - National Health and Nutrition Examination Survey Homepage (cdc.gov)
		EU-Statistics on Income and Living Conditions (EU-SILC)	EU Statistics on Income and Living Conditions • European University Institute (eui.eu)
		Scottish Health Surveys (SHeS)	Scottish Health Survey - gov.scot (www.gov.scot)
		The Irish Longitudinal study on Ageing (TILDA)	https://tilda.tcd.ie/
		Canadian Community Health Survey (CCHS)	Canadian Community Health Survey - Canada.ca
Health insurance claims data	Include demographic information, date of service, providers, type of service, healthcare expenditures, procedures, (diagnoses).	Belgian compulsory health insurance (BCHI) data	Données de santé (ima-aim.be)
		Ontario Health Insurance Plan (OHIP) claims,	https://www.health.gov.on.ca/en/public/programs/ohip/
		The German Statutory health insurance data	Statutory health insurance - GKV-Spitzenverband
Hospital morbidity data	Hospital separation data (discharges, transfers and deaths) from all non-psychiatric hospitals. Include relevant clinical data (e.g. primary and secondary diagnosis), procedures performed, length of stay, residential	Minimal Hospital Data (MHD), Belgium	https://www.health.belgium.be/en/node/23774
		Hospital morbidity data system, Western Australia	Hopsital Morbidity Data System (HMDS) Reference Manual - Part A Data Element Defintions (health.wa.gov.au)
		Hospital Episodes Statistics (HES), UK	Hospital Episode Statistics (HES) - NHS Digital

	address and demographic characteristics of patients.		
		Discharge Abstract Database (DAD), Canada	https://www.cihi.ca/en/discharge-abstract-database-metadata-dad
Medical records data	Primary care data. The data set includes demographic information, date of death, age of deceased, cause of death, occupation of deceased and the health authority to which a person is registered.	Intego data, Flanders (Belgium)	https://www.intego.be/fr/
		National Health Service Central Register (NHS-CR), UK	National Health Service Central Register | National Records of Scotland (nrscotland.gov.uk)
		Medicare & Medicaid Services (CMS) data, USA	https://www.cms.gov/research-statistics-data-and-systems/research/mcbs?redirect=/mcbs/
Disease registries	All incident cases. Include patient-level demographics data, event dates, cancer site, cancer morphology, first course of treatment (surgery, radiation).	Belgian cancers registry	https://kankerregister.org/default.aspx?lang=EN
		Dutch Pediatric and Adult Registry of Diabetes; DPARD, the Netherlands	https://dica.nl/dpard/home
		Diabetes-Patienten-Verlaufsdokumentation; DPV, Germany	https://buster.zibmt.uni-ulm.de/projekte/DPV/
		Norwegian Surveillance System for Communicable Diseases (MSIS)	Norwegian Surveillance System for Communicable Diseases (MSIS) (helsedata.no)
Prescription drugs data	Pharmacy dispensing records. Contains information on all prescribed drugs dispensed under the medical card scheme. The data include patient's age; gender; and for each medicine dispensed, the non-proprietary drug name, proprietary drug name, strength, and quantity dispensed. All prescription items are coded using the WHO ATC classification system.	Pharmanet data, Belgium	https://metadata.ima-aim.be/fr/app/bdds/Fu
		Ontario Drug Benefit (ODB) Claims, Canada	https://data.ontario.ca/dataset/ontario-drug-benefit-odb-database
		Scottish NHS prescriptions data	https://www.opendata.nhs.scot/dataset/prescriptions-in-the-community
Census data	Focus was on demographic and socio-economic information in a comprehensive and detailed way. Could also include subjective information and opinions of individuals, such as their perceived health and the quality of their environment. Could include characteristics of geographic unit	Belgian census data	https://census2011.fgov.be/
		US census data	https://www.census.gov/data.html

	(e.g. median household income, racial composition, employment rates).		
Labour market and social protection data	Holds detailed information about the social security benefits and tax credits received by each individual. Contains daily information on employment histories, information on transfer payments and wages, number of unemployed, number of individuals receiving benefits, full demographics, profession, etc.	Integrated Employment Biographies of the Institute for Employment Research (IAB), Germany	https://fdz.iab.de/en/our-data-products/individual-and-household-data/siab/
		The Crossroads Bank for Social Security, Belgium	https://www.ccc-ggc.brussels/en/observatbru/data-sources/crossroads-bank-social-security-datawarehouse-labour-market-and-social
Vital statistics (birth and death records)	Population births and deaths registration. Contains full demographics, mother's and father's details, date of death, underlying causes of death, other conditions present	Vital registrations from Statbel, Belgium	https://statbel.fgov.be/en
		National Death Index, USA	https://www.cdc.gov/nchs/ndi/index.htm
Environmental data	Concentrations of air pollutants	European Environment and Epidemiology (E3) Network: E3 geoportal environmental datasets	ECDC Geoportal | E3 Network (europa.eu)
		Belgian Interregional Environment Agency (IRCEL - CELINE) data	https://www.irceline.be/en
Provider / institutional files	Data collected at provider or organisation level. Provides resource information (e.g. number of physicians, specialists, hospitals per 100,000 residents), global information on indicators of medical consumption	National institute for health and disability insurance (NIHDI), Belgium	https://www.inami.fgov.be/fr/Pages/default.aspx
		American Medical Association Physician Masterfile, USA	https://www.ama-assn.org/practice-management/masterfile/ama-physician-masterfile
		American Hospital Association, USA	https://www.aha.org/

2.3. TYPES OF DATA LINKAGE: AD HOC VS SYSTEMATIC DATA LINKAGE

Data linkages can be ad hoc (project-based), or routine (systematic) data linkage. An ad hoc data linkage focuses on just one or a small number of research projects and in this form it is widely practised worldwide in clinical, health services, and public health research. Often it links records of harmful exposures or beneficial healthcare interventions with records of health outcomes. In contrast, systematic data linkage is undertaken on a proactive and systematic basis for health data pertaining to an entire population and it is aimed to be used as data infrastructure for an indefinite number of future (and as yet undefined) research projects. Systematic data linkage involves the maintenance of a permanent and continuously updated master linkage file and a master "statistical linkage key (SLK)) (27). The contrasting characteristics of these approaches are shown in Table 2.2 below.

Table 2.2: Main differences between ad hoc and systematic data linkage

	Ad hoc data linkage	Systematic data linkage
Purpose	Supports one research project (or a small number) with known objectives.	Supports an indefinite number of mostly unknown future research projects.
Data sets	Limited to those needed for the known research objectives (often 2-3 data sets).	Unlimited – the more data sets the more versatile and effective is the system.
Data requirements	Usually, partial identifiers and clinical data come together.	Only requires partial identifiers – clinical data can be sought later on a project-by-project basis
Time of activity	Data linkage activity closes once links between the specified data are in place.	Open-ended and requires continuous updates of the links as new data arrive.
Storage of links	Usually, links are stored as an integral part of the research project data	Requires a dedicated storage mechanism for links = the master linkage key.
Funding	Usually draws on the research grant used to fund the project.	Requires dedicated infrastructure funding for a central, ongoing unit

Source: Tom Briffa and Jane Heyworth. Introductory analysis of linked health data course. Principles and hands-on applications. Version 3.5s February 2019 (28).

2.4. DATA LINKAGE METHODS

The linkage of two or more data sources requires at least one common identifier between these data sources. Data linkage is a relatively straightforward process in situations where perfect unique personal identifiers exist in all the datasets to be linked or where identifying information is recorded without error.

In such circumstances, the matching process can be limited to a simple sort and merge of the data sources by personal identifiers. However, perfect datasets are rare, and it is more common that there will be discrepancies in identifying information between pairs of records belonging to the same person. In these situations, exact matching using these personal identifiers miss a significant number of true links. Depending on the quality of the data and the availability of unique identifiers, one can distinguish overall three broad approaches to linkage methods: deterministic (rule-based) methods, probabilistic (score-based) methods and a group of newer approaches such as techniques that make use of advanced machine learning algorithms (29,30). Although these methods are usually treated as distinct methods, in practice, linkage studies often use a combination of deterministic and probabilistic methods, using initial deterministic steps to reduce the number of comparison pairs for subsequent probabilistic linkage (31,32).

2.4.1. Deterministic linkage methods

Deterministic linkage or rule-based methods are relatively straightforward approaches to linkage. Deterministic algorithms indicate whether record pairs agree or disagree on a given set of identifiers, where agreement on a given identifier is assessed as a discrete “all-or-nothing” outcome (20). These methods, which typically require exact agreement on identifiers, are useful when there are unique identifiers or a set of several attributes that are highly discriminative, completed and accurate. In this ideal situation when unique and reliable identifiers exist, the linkage process can be reduced to a simple sort of the records by unique identifiers (21). This one-step procedure using a single unique identifier or a set of personal several discriminative attributes is called “exact” deterministic linkage (20,33). The usual common unique identifiers include for example the Social Security Number (SSN), the Health Insurance Claim Number (HICN) and the Medical Record Numbers in USA, the

national insurance number in UK, the community health index in Scotland (6,29,34– 37) and the national register number in Belgium (38–40).

However, in most situations, there are no such unique identifiers and only common non-unique identifiers like names, sex, date of birth, race, postcode exist between datasets to be linked. For such situations, exact matching using these personal identifiers miss a significant number of true links (33). Therefore, the strict deterministic approaches are usually relaxed by allowing a linkage on a specified set of partial identifiers (e.g. surname, sex and postcode) (6,29,34–36) allowing small differences in identifiers and using a succession of rules. In other words, in case of partial identifiers, the deterministic linkage approaches make use of a pre-determined set of rules that will be executed in a particular order for classifying pairs of records as belonging to the same individual. So a step-wise algorithmic linkage involving a series of progressively less restrictive steps to allow variation between record attributes (also referred as “multiple-step strategy” or “iterative” deterministic linkage) is used. A record pair is classified as “linked” if it meets the criteria or parameters at any step; otherwise, it is classified as “non-linked” (20). For example, the three-step deterministic algorithm used in England to link hospital admission records for the same individual in Hospital Episode Statistics is based on a sequential set of rules looking for agreement on a combination of identifiers (41,42): 1. NHS number, date of birth and sex; 2. Local patient identifier, hospital provider, date of birth, sex and postcode; 3. Date of birth, sex and postcode. The steps are continued until as many records as possible are correctly linked (19).

Deterministic methods are designed to avoid false matches (i.e. records from different individuals link erroneously), since it is unlikely that different individuals will share the same set of identifiers, although this can occur where there are identifier errors. On the other hand, deterministic methods requiring exact agreement on identifiers are prone to missed matches (i.e. records from the same individual fail to link), as any recording errors or missing values can prevent identifier agreement (6,42,43). A deterministic linkage method is most applicable when the number of records to be matched is relatively small, there are a limited number of data attributes for linkage, and there are minimal recording errors within the underlying datasets, i.e. files with high-quality data (35). When the number of data attributes and rules required is small, the development of the deterministic matching algorithms is relatively simple and is

easy to implement. The more the linkage involves large datasets with complex characteristics, the more complicated the rules-based matching routines become (44). Deterministic matching systems are typically less sensitive to errors/discrepancies in the data and as a result will miss more links. In most administrative data collection systems, the datasets are large, increasing the potential for duplicates, human error and discrepancies. The system design must allow for complex error patterns within true links, enabling the determining of links within and between data files. Furthermore, the deterministic approach ignores the fact that certain identifiers or certain values have more discriminatory power than others (20).

2.4.2. Probabilistic linkage methods

Probabilistic methods, also known as the Fellegi–Sunter algorithms (45), were proposed as a means to overcome some of the limitations of deterministic linkage, and to allow linkage in the presence of recording errors and/or without using a unique identifier (6,20). Probabilistic strategies take advantage of differences in the discriminatory power of each attribute and involve the calculation of similarity scores (match weight), as well as decision rules, to classify record pairs as linked, potentially linked (treated as dubious records in most linkage tools) and non-linked (9,20,33,45). It can also deal with some inconsistencies between records with missing data, i.e. it has the capacity to link records with errors in the linking fields (20).

Probabilistic linkage approaches dominate in traditional record linkage applications and remain an effective and efficient way to solve the record linkage problem today (46). Newcombe was the first to propose probabilistic methods, suggesting that a match weight could be created to represent the likelihood that two records are a true match, given agreement or disagreement on a set of partial identifiers (9). Fellegi and Sunter (1969) formalised mathematical methods for considering a record ‘‘linked.’’ Their seminal work defined a clear linkage rule that assigns a probability that two records from separate files represent the same person (or entity) (45). However, the Fellegi and Sunter algorithm has been criticised because of his accuracy and efficiency (7). Methods have since been developed that improve the accuracy and efficiency of Fellegi and Sunter’s original work, such as the methods proposed by Winkler (1993) (47) and Jaro (1995) (48). Probabilistic linkage requires investment in software that will do the match.

One of the concerns in probabilistic approaches is the comparison space which represents the Cartesian product made up of all possible record pairs in files to be linked. When dealing with large files (e.g. administrative claims files), considering the entire Cartesian product is often computationally impractical. In these situations, it is advisable to reduce the comparison space to only those matched pairs that meet certain basic criteria. Blocking strategies are used to reduce the set of potential matches to a more manageable number (20). Matched pair identified in the blocking phase are compared on each linkage identifier, producing an agreement pattern. The weight assigned to agreement or disagreement on each identifier is assessed as a likelihood ratio, comparing the probability that true matches agree on the identifier ("m-probability") to the probability that false matches randomly agree on the identifier ("u-probability"). When two records agree on an identifier, an agreement weight is calculated by dividing the m-probability by the u-probability and taking the log2 of the quotient. When two records disagree on an identifier, a disagreement weight is calculated by dividing 1 minus the m-probability by 1 minus the u-probability (20).

Sometimes, partial agreement weights for string comparators are assigned in situations where two strings do not match character for character can account for minor typographical errors, including spelling errors in names, addresses or transposed digits in dates or SSNs (49,50). Partial agreement weights for string comparators can account for both the length of the string and common human errors made in alphanumeric strings. If all of the characters in a string are matched character by character across two files, then the agreement weight is maximised (set at the full agreement weight). If there are no characters in common, then the agreement weight is zero (50). The full agreement weight for the identifier can then be multiplied by the string comparator value to generate a partial agreement weight. For example, if the full agreement weight for the first name is 12 and the string comparator value is 0.95, then the partial agreement weight for the match between the first name on one record and the first name in another record would be equal to 12*0.95, or 11.4. Once the weights, full and partial, for each identifier have been calculated, the linkage score for each matched pair is equal to the sum of the weights across all linkage identifiers. Use of string comparator methods may significantly improve match rates if a large number of typographical errors are expected.

An initial assessment of the linkage quality can be gained by plotting the match scores in a histogram. If the linkage algorithm is working properly, then the plot should show a bimodal distribution of scores, with one large peak among the lower scores for the large proportion of likely non-matches and a second smaller peak among the higher scores for the smaller set of likely matches. The cutoff threshold for match/non-match status will be a score somewhere in the trough between the largest and smallest peaks. Depending on the research question and the nature of the study, the initial threshold can be adjusted to be more conservative (higher score) or more liberal (lower score). A more conservative threshold will maximise the specificity of the linkage decision, as only those record pairs with a high score will be counted as matches. Conversely, a more liberal threshold will maximise the sensitivity of the linkage decision to possible matches (20).

In summary, the probabilistic linkage approaches require the following steps (20):

- Estimate the m and u probabilities for each linking variable using the observed frequency of agreement and disagreement patterns among all pairs, commonly generated using the expectation-maximisation (EM) algorithm described by Fellegi-Sunter
- Calculate agreement and disagreement weights using the m and u probabilities
- Calculate a total linking weight for each pair by summing the individual linking weights for each linkage variable
- Compare the total linkage weight to a threshold above which pairs are considered a link. The threshold is set using information generated in step 1.

2.4.3. Alternative data linkage methods

Deterministic and probabilistic methods are the most commonly used linkage approaches. However other methods are available for researchers who have more challenging linkage scenarios. The EM algorithm is an iterative approach that can be used for estimating the weights (m- and u-probabilities) under less restrictive assumptions and provides very accurate estimates of m- and u-probabilities in situations where the amount of typographical errors in the identifiers is minimal, but

performs poorly when the identifiers contain numerous typographical errors (49,51). EM algorithm improves computational procedures in applications of the Fellegi-Sunter model of data linkage (51). In addition, the Bayesian approach (52) is also an alternative approach to the frequentist approach.

In recent years, and as a result of new advances, machine-learning approaches (53) have been applied to record linkage (46). Indeed, alternative methods for supervised classification methods, such as logistic regression, Support Vector Machines (SVM), Random Forests and Gradient Boosting, and unsupervised classification methods (with or without training data, respectively), have found their way into the domain of data linkage (18). While supervised techniques typically classify each record pair individually, so-called `collective' linkage techniques consider whole clusters of linked records (such as several individuals living at the same address) with the aim to find an overall optimal and consistent linkage solution for an entire database (54). Unsupervised machine-learning techniques, on the other hand, are mostly employed in linkage situations where multiple records of the same individual might exist (for example all hospital records for the same patient), or where records from groups of individuals need to be linked (such as all babies born to the same parents) (55–58). In contrast with the standard Fellegi–Sunter application which uses indexing and blocking, machine learning-based approaches are likely to use the more sophisticated clustering approach to indexing. Indexing may also use network information to include, for example, records for individuals that have a similar place in a social graph. When linking lists of researchers, one might specify that comparisons should be made between records that share the same address, have patents in the same patent class, or have overlapping sets of coinventors. These approaches are known as semantic blocking, and the computational requirements are similar to standard blocking (18).

2.5. CHALLENGES AND PRIVACY CONCERNS

Although linking administrative data to survey data may be perceived as a relatively economical and straightforward way to enhance survey data, the actual process can be costly, time-consuming and challenging. Traditionally, the challenges inherent in linking administrative data with survey data sources can be grouped into technical challenges and legal, ethical constraints. This section briefly discusses the main

challenges and privacy concerns arising when linking administrative data and survey data.

2.5.1. Technical challenges

The technical challenges inherent in linking survey data with administrative data are mainly related to the data quality on one hand and to the linkage errors on the other hand (59).

Challenges due to data quality

Administrative data are primarily not designed for epidemiological research nor for linkages. They are subject to missing data in case of incomplete recording or when a person fails to interact with a service and is therefore not captured in the administrative data. Data linkage adds a further dimension: missing or inaccurate data can also be introduced if the individual's record could not be accurately linked due to insufficient identifying information (23).

Linking survey data with administrative data requires that the two data sources contain overlapping information, i.e. at least one common variable. The most straightforward situation occurs when both data sources contain a unique personal identifier, such as a social security number. In this ideal case, the data can be directly linked, usually with almost no errors. However, such situations are rather rare. In the absence of unique identifiers, combinations of other available individual characteristics must be used instead, such as name, sex, address, and date and place of birth, to identify identical subjects in both data sets. In this case, data linkage become challenging since these are "imperfect identifiers" as they may not be unique, and they may vary over time (addresses and names change). Sometimes one or the other data set will contain typos or other logging errors, or inputs may be missing entirely. Linking two data sources based on imperfect identifiers is thus not straightforward; it requires a multi-step, iterative process (60,61) that can be time- and resource-intensive. Therefore, depending on the data quality, the appropriate linkage methods should be chosen (see section above 2.4).

Challenges due to linkage errors

The appropriate use of linked administrative data for research poses particular challenges, for example with regard to bias because records cannot be linked or are incorrectly linked (23). Linkage errors may occur if records are incorrectly linked (false matches) or when the same person fails to be linked (missed matches) (6,20) (see Table 2.3). Linkage errors typically occur where there is no unique identifier across data sources (62) and could lead to biased results and requires appropriate analysis methods (14,59,63,64).

Other operational challenges

Another challenge that researchers face in data linkage is the principle of proportionality. According to the principle of proportionality, or "data minimisation", the least amount of personal data should be processed that is necessary to achieve the purpose (65). The selection of the minimum required number of adequate, relevant variables must be done precisely before the linkage process. However, it is often difficult to decide on the variables to be involved in the linkage process. The more information there is in both data sources, the more difficult this task becomes.

Another consideration for researchers wishing to link data is the infrastructure needed to store and access the linked data.

Finally, analysing linked datasets raises a number of statistical challenges for researchers. Although linked data have several advantages, it is important to keep in mind that the limitations of both data sources remain even after the linkage. In addition, in case of linkage errors, specific statistical methods need to be applied (6,24).

2.5.2. Legal challenges and privacy concerns

Institutional, ethical approvals processes

The main challenge when linking survey and administrative data is to deal with privacy and confidentiality issues (66) and the data restrictions resulting from them, especially with the implementation of the GDPR in 2018. Because of confidentiality issues, institutional or ethical review boards (IRB) approval is often required to access and

link administrative data (25). However, such IRB approval process is usually complex and time-consuming, especially when the linkage is not consent-based.

Furthermore, several legal, ethical and cultural considerations may significantly constrain the extent to which researchers can link data in practice. These may include variations and uncertainties over what is permissible, questions around consent, and concerns over public acceptability and trust (67).

Respondents' rights and trust - opt-in informed consent

Privacy concerns are justified and necessary, as information in administrative data are collected as part of administrative processes that are usually conducted without the explicit agreement of the individuals involved. That means that the individuals whose information is collected never consented to the use of their administrative records for scientific research. The respect of respondents' rights and the duty to maintain their trust are also very important. According to the new EU data Act, trust and altruism are essential in secondary data use (68). When researchers plan to link data as part of a future survey, citizens must be able to decide whether they want to share their data, be informed that their data is being used and who is using them. The easiest way to deal with this is to inform survey respondents about the intention to link survey and administrative data, along with any associated risks, and to ask for permission to use the collected survey data in such a way through opt-in informed consent (69–72). However, obtaining the opt-in linkage consent from all respondents is a challenging task. In addition, such an approach could introduce consent bias if individuals who give permission to link survey and administrative data sets likely differ systematically from individuals who deny consent (69). To increase consent rates and reduce potential consent bias, some authors argue that consent for linkage should be sought at the beginning of the survey rather than at the end (questions at the beginning of a survey obtain higher consent rates) (69).

To link historical survey data to administrative data, there are exceptions to the need for informed consent, especially, when it is impossible or unreasonable to contact the study participants (71,72). The GDPR contains specific exemptions to informed consent as a legal basis for the use of data to escape a 'consent or anonymised approach' or a 'fetishisation of consent', especially in the case of observational health research (73). Some countries have legislated legal exemptions to consent for data use for scientific purposes, which creates legal space for linking data even where an

individual has not explicitly consented to the linkage of the data. An example of such exemptions would be in case of ad hoc linkage when the linkage between administrative data and survey data was not foreseen ahead at the time of survey implementation. Such legal exemptions usually require demonstrating that the importance of the scientific inquiry for which linked data are being requested outweighs any related privacy concerns. However, this is not always sufficient to obtain the authorisation to access and link the data.

Privacy preserving and separation principle

For confidentiality and other reasons, the separation of data linkage processes and analysis of linked data is generally regarded as best practice. However, the 'black box' of data linkage can make it difficult for researchers to judge the reliability of the resulting linked data for their required purposes (23,24). To preserve privacy and avoid disclosing sensitive information, data linkage often relies on the separation principle of linkage and analysis processes, meaning that those conducting the linkage (often through trust third party (TTP)) only have access to a set of identifiers, whilst those analysing the linked data only have access to de-identified attribute data (23). However, such an approach may cause an important delay in the linkage process because of administrative steps that take time (e.g. signature of official agreement between involved parties). Furthermore, although this approach reduces the risk of disclosing sensitive information about individuals, it implies that important aspects of the linkage process are obscured which makes it difficult for researchers to judge the reliability of the resulting linked data for their required purposes (23,62,74).

In addition, the high risk of identifying an individual's personal details in the linked data set usually means that a high degree of data anonymisation is required, which severely restricts access to the data and reduces its research potential. Therefore, the manager of the administrative data (typically a government institution) usually requires the exclusion or aggregation of information, to mask its identifiable personal properties, and may further restrict data access to protect individuals against data misuse. There are trade-offs between preserving the research potential of the data and controlling data availability that need to be resolved through compromise (61).

2.6. EVALUATING LINKAGE QUALITY

The assessment of the quality of the linkage is crucial to detecting possible errors and to take into account the limitations of the linked data in the statistical analyses. Several methods can be used to evaluate the quality of data linkage (14,29). These methods focus on identifying potential sources of bias (that is, which characteristics are associated with errors) by examining the characteristics of records that are linked versus those that are unlinked, or that have high versus low quality identifier data, or that are easily identifiable as having been linked incorrectly (e.g. through quality control checks) (75). Linkage quality is generally described in terms of the types of linkage error and the magnitude of these errors (30). Achieving high linkage quality is essential for ensuring and maintaining the quality and integrity of research and related outputs based on linked data.

2.6.1. Linkage error

Linking survey data with administrative data usually relies on imperfect identifiers because of the lack of a unique personal identifier. However, the quality of record linkage is reliant upon the availability and accuracy of common identifying variables. Imperfect identifiers are not sufficiently discriminative, prone to missing values, recording errors, and change over time (24). Irrespective of the linkage methods implemented, use of imperfect and dynamic identifiers can lead to linkage error (6,14,76). Linkage errors arise when pairs of records are incorrectly classified and manifest themselves as false matches (also called false positives) or missed matches (also called false negatives). False matches occur when records from different individuals link erroneously; while missed matches occur when records from the same individual fail to link (23). False matches and missed matches are increasingly being recognised as a potential source of bias in results from studies using linked data (55,77). However, it may be difficult for users of linked data who have not been involved in the actual linkage process to assess the extent to which they influence the results (24). Analogous to false positives and false negatives, false matches or missed matches can be viewed through a diagnostic accuracy lens (see Table 2.3). Linkage procedures frequently involve managing trade-offs between false matches and missed matches because reducing false matches will tend to increase the risk of missed matches, and vice versa (30).

Table 2.3: Linkage accuracy tool

Assigned link status	Match status		
	Match (same individual)	**Non-match (different individuals)**	
Link	True match (a)	False match (b)	Total links (a+b)
Non-link	Missed match (c)	True non-match (d)	Total non-links (c+d)
	Total matches (a+c)	Total non-matches (b+d)	Total records pairs (a+b+c+d)

Source: adapted from Harron, Katie (2022): "Data linkage in medical research." (14)

2.6.2. Impact of linkage error on research outcomes

Linkage error can threaten the reliability of results based on analyses of linked data. Errors in linkages involving administrative data are often unavoidable, specifically when imperfect identifiers are used for linkage. The impact of linkage error on analysis of linked data depends on the structure of the data, the distribution of error, and the proposed analysis. With health data, the number of false matches and missed matches can directly affect the estimation of prevalence or incidence rates as well as the associations. The impact of linkage bias can be high even when the error is small, as small amounts of linkage error can result in substantially biased results (78). Conversely, a large amount of error will not necessarily produce bias. This is because the impact of linkage error depends more on how it alters the structure of the data than on the number of errors that have occurred. For example, if an event is rare, it would require only a small decrease in specificity for many or the majority of assigned events to be false, with consequent implications for any conclusions drawn from the data (30,78).

False matches (low specificity) lead to overestimates of prevalence. For instance, when a record is linked but no link should have been made (e.g. linking a survivor to a mortality record), this can have implications for prevalence estimates (such as overestimating a rate). False matches are a further challenge. Irrespective of the levels of linkage errors, false matches reduce the magnitude of the association. When records from two different individuals are linked together, false matches can add noise

to estimates, dilute true relationships, and tend to lead to bias towards the null hypothesis, i.e. they can increase the likelihood of a type 2 error (55). If false matches depend on individual characteristics (e.g. sex, because of maiden/married name inconsistencies) this may lead to biased estimates of association, e.g. if sex is related to both the exposure and outcome of interest (23).

Missed matches lead to underestimates of the prevalence. When unlinked records are excluded from analyses, one consequence is a reduced sample size and statistical power and, irrespective of the levels of linkage errors, reduce precision. If linkage is 'informative' (e.g. linkage to a disease register indicating the presence of a particular condition), a consequence of missed matches can be under-ascertainment of exposures or outcomes. However, with the large sample sizes available in administrative data, a more serious problem associated with missed matches is selection bias, which occurs when particular groups are systematically less likely to link (non-random or differential linkage error) (64,79) and hence are excluded from analysis. Systematic reviews of studies comparing the characteristics of linked and unlinked records have identified that more vulnerable or hard-to-reach populations are often missed, with the probability of a missed match being associated with a range of characteristics including sex, age, ethnicity, deprivation and health status (64,80). Consequently, the linked data may not be representative of the population of interest, which can reduce the study's external validity (loss of generalisability) or may not capture subgroups that are of particular interest. As these demographic variables are often associated with exposures or outcomes of interest, differential rates of linkage error may also introduce bias. For example, unlinked mortality records in one particular ethnic group could lead to a distorted comparison of mortality rates by ethnicity (81). If unlinked records are to be excluded from analysis, selection bias (or collider bias) can occur if selection into the linked dataset is related to both an exposure and an outcome of interest (82). For example, suppose it is more difficult to link records for low birthweight babies and also more difficult to link records from mothers who smoke. In this case, records for low birthweight babies that are successfully linked are more likely to be from mothers who do not smoke (since, in this example, records from mothers who smoke are more difficult to link). Conditioning on linked records could therefore induce a protective relationship between maternal

smoking and low birthweight, analogous to the birthweight paradox described in epidemiological literature (24,83).

Therefore, adjusting for these biases could provide more robust results using data with considerable linkage errors. Studies based on high-quality linked data in developed countries show that even minor linkage errors, which occur when records of two different individuals are erroneously linked or when records belonging to the same individual are not linked, can impact bias and precision of subsequent analyses. The authors evaluated the impact of linkage quality on inferences drawn from analyses using data with substantial linkage errors in rural Tanzania (84).

Table 2.4 gives an overview of the types of linkage error and their impact on the results

Table 2.4: Types of linkage error and how they arise

Error type	**False links**	**Missed links**
Also known as	False positives	False negatives
What is the error?	Records are linked but they belong to different individuals	Records from the same individual are not linked
Common sources of error	Identifiers do not discriminate well between individuals: • Large file sizes • Many people share identifiers e.g. age and sex	Usually from errors in identifiers: • Typographical errors • Changes over time (e.g. married women changing their surnames) • Missing or invalid data
Type of bias that might result	Information bias (i.e. misclassification or measurement error)	Selection bias or information bias

Source: Amanda Kvalsvig et al. (2019). "Linkage error and linkage bias: A guide for IDI users" (30).

2.6.3. Measuring linkage quality

Typically, researchers become aware of errors when an invalid or implausible combination is found following early applications of consistency or logic checks. For example, a hospital record of a full-term delivery occurring after a hospital record indicating a hysterectomy (84), or hospital records indicating that a person was hospitalised after their date of death (30).

Assessing linkage quality is vital and allows identifying limitations of the linked data to be considered within analysis. Furthermore, such evaluation will improve the quality

and transparency of epidemiological and clinical research using the linked data (24). Evaluation of linkage quality is typically done either through systematic quality assessment within large-scale linkage systems or on a project-specific basis, and can be done by the data linker, the data-user, or a combination of both. For large-scale linkage systems, systematic quality assessment might include regular consistency checks and manual review of linked and unlinked records. For project-specific linkages, the nature of evaluation of linkage will depend on the nature of the planned analyses and the information available. For example, a particular study question might require high specificity, in which case evaluation would focus on the false match rate.

Errors that occur during the linkage process (false matches and missed matches) can lead to biased results, although the extent of this bias in research based on linked data is difficult to measure, as standard measures of linkage error do not necessarily allow understanding of the impact of these linkage errors on results (85).

Measures of linkage error typically reported in the literature include match rate, sensitivity, specificity, positive predictive value and negative predictive value (6,21,23,24,86,87). In practice, the number of non-matches will usually far outweigh the number of matches, therefore the positive predictive value and sensitivity are more informative than the specificity and negative predictive value. F-measure, the harmonic mean of positive predictive value and sensitivity is commonly used to represent the quality of a linkage with just one number (20,21). Using a single metric makes it easier to compare linkages. The use of the harmonic mean results in higher scores only when both precision and recall have higher scores, unlike a simple average. Table 2.5 presents these measures and how to calculate them.

Table 2.5: Standard metrics of linkage errors

Measure	Definition	Formulae
Match rate	Proportion of records that were linked	(a+b)/(a+b+c+d)*
Sensitivity (or Recall)	Proportion of matches that are correctly identified as links	a/(a+c)
Specificity	Proportion of non matches that are correctly identified as non-links	d/(b+d)
Positive predictive value (or Precision)	Proportion of detected links that were true	a/(a+b)
Negative predictive value	Proportion of non-matches that are not true links	d/(c+d)).
F-measure	The harmonic mean of precision and recall	$\frac{2 \text{ X Precision X Recall}}{\text{Precision} + \text{Recall}}$

** a, b, c and d refer to Table 2.3.*

The most appropriate linkage quality measures depend on the purpose of the linkage and the end use of the linked data: avoiding false matches is important for some studies, whereas for others, a high match rate may be more desirable. For example, consider linkage between a cohort dataset and a cancer registry. A highly specific linkage (i.e. one where there were few false matches) would mean that all participants identified as having cancer really did have the disease. However, a strict linkage strategy may prevent some links from being identified, meaning that some of the controls also had cancer, but had not been identified. This could lead to dilution of any true associations and would mean that the linked data may not be useful for providing estimates of cancer incidence. Conversely, if a more sensitive linkage were achieved, incidence estimates would be more accurate, as more cancer cases have been identified. However, some of the records may be falsely linked, meaning that a number of controls are misclassified as cases. It is important to understand the implications of linkage errors when considering study design and analyses (6).

A major limitation of the standard measures of linkage error is that they do not provide information on how results of analyses might be affected in terms of bias (24,85) and are not always relevant or interpreting them is not always straightforward. For

example, match rate is only helpful if you know how many records from a particular dataset should be linked (6).

Several approaches to evaluating linkage quality have been proposed to overcome limitations of standard methods. The use of these methods can help researchers using linked data to understand the potential impact of linkage error on results, and comprise (6,23,24):

- Comparing linked data with reference or 'gold-standard' datasets where the true match status is known
- Structured sensitivity analyses where a number of linked datasets are produced using different linkage criteria
- Comparisons of characteristics of linked and unlinked data to identify any potential sources of bias
- Statistical methods accounting for linkage uncertainty within analysis (e.g. using missing data methods), or using population weights to account for groups or people who are more or less likely to be linked
- Quality control checks (implausible scenarios).

Table 2.6 summarises these main approaches to evaluating linkage quality.

Table 2.6: Summary of approaches to evaluating linkage quality

Approaches	Purpose	Strengths	Limitations	Technical requirements
'Gold standard' or reference data	To quantify errors (missed matches and false matches)	Easily interpretable; allows linkage error to be fully measured	Representative gold standard data are rarely available	A representative group of records for which true match status is known; data linker capacity to perform evaluation (researchers rarely have access to gold standard data)
Comparing characteristics of linked and unlinked data	To identify subgroups of records that are more prone to linkage error and are potential sources of bias	Straightforward to implement and easily interpretable	Cannot be applied if systematic differences are expected between linked unlinked records (e.g. if linking to death register)	A linkage design where all records in at least one file are expected to link: provision of record-level or aggregate characteristics of unlinked records to researchers Where not all records are expected to link (e.g. linkage between a study population and a disease registry), comparisons may need to be performed on a higher level. For example, age and sex distributions of linked records could be compared with distributions in population data, to establish how representative the linked data are of the target population, i.e. to explore any evidence of selection bias
Sensitivity analyses	Assesses the extent to which results of interest may vary depending on different levels of error, and the direction of likely bias	Straightforward to implement	Results may be difficult to interpret as false matches and missed matches may impact on results in opposing or compounding ways	Provision of information on the strength of the match (e.g. deterministic rule or probabilistic match weight)
Post-linkage data validation	To estimate minimum false-match rates by identifying implausible scenarios within the data.			

Source: Harron et al. (2017) - A guide to evaluating linkage quality for the analysis of linked data (24); Harron et al. (2017) - Challenges in administrative data linkage for research (23); Harron et al. (2016) - An introduction to data linkage (6)

2.6.4. Addressing linkage error in analysis of linked data

Mitigating the impact of linkage error is essential to deal with linkage bias. The goal of adjusting for linkage bias is to produce an analysis output (e.g. estimate of effect) that is closer to the true value than the unadjusted (biased) result (88). When linkage error is identified as a possible source of bias, methods to adjust for these biases should be used, which can help provide more robust results (24). Evaluation of linkage quality can guide decisions about appropriate study design. For example, if linkage is used to identify individuals with a particular condition or disease (informative linkage), high levels of missed matches will lead to under-ascertainment, meaning that cohort study designs may be unsuitable (particularly for deriving estimates of prevalence or incidence). Where linkage rates are too low, researchers may conclude that linked data are not fit for these purposes. On the other hand, a case-control study may still be valid, whereby a high threshold is used to identify cases and a low threshold is used to identify controls (assuming no other biases are present). In this scenario, records for which there is uncertainty about linkage would not be included in analysis (24). Accounting for linkage error in analysis is an ongoing area of methodological research (14) but includes approaches that view uncertainty in linkage as a missing data problem best handled with some form of multiple imputation or weighting, and those that attempt to quantify and adjust for errors using quantitative bias analysis (63). Overall, techniques for addressing linkage error can be broadly grouped into probabilistic analysis, sensitivity analysis and bias analysis (29).

Probabilistic analysis: including multiple imputation and inverse probability weighting.

If individual-level information about matching status (correct or incorrect) is available, then match probabilities can be inputted into multiple imputation to handle missing values due to unlinked or equivocal records (85,89). Furthermore, information from match weights can be incorporated into imputation procedures, making use of variable distributions in candidate links (known as ‘prior-informed imputation’ (PII)). PII is a more flexible method for dealing with linkage uncertainty. The method incorporates information from ‘auxiliary’ variables, such as individual characteristics associated with linkage quality to help correct for selection biases without requiring

identifiers (85). PII aims to select the correct value for variables of interest. Unlike existing methods which accept a single complete record as link, PII, allows more than one candidate linking record to be considered in analysis. PII can be utilised for data that belong to records that cannot be matched unequivocally. In this way, the information from all potential matches is transferred through to the analysis stage. This procedure allows for the propagation of matching uncertainty through a full modelling process that preserves the data structure. Standard multiple imputation has also produced unbiased and efficient parameter estimates in simulation studies (85,89).

In the inverse probability weighting models, the analysis can be weighted to take error into account. Such an approach has been successfully used in previous studies (90– 92). Probabilistic analysis requires access to uncertain links and estimates of match probabilities that may be hard to estimate. A second limitation of these techniques (one that is likely to be addressed in future development) is the complications that arise when the unit of analysis is affected by clustering (e.g. when two records are counted as one person if linked and two people if not) (29).

Sensitivity analysis: in which the analyst varies the threshold for accepting record pairs as links, moving up or down the spectrum of agreement. If direct adjustment is not possible but record-level linkage weights are available, researchers can gain some indication of the likelihood of differential linkage error using a sensitivity analysis approach, i.e. by repeating an analysis using different cut-offs to understand how sensitive the analysis results are to differing cut-offs; this approach can generate insights about linkage error by examining how the results change as the balance shifts between false positives and false negatives (30). For example, Lariscy et al (2011) utilises sensitivity analyses to examine how ethnic mortality differentials change with modification of the National Center for Health Statistics (NCHS) recommended match score cut-off points for death ascertainment (81). While sensitivity analysis is probably the most common example of analysis accounting for linkage error to date, it is also limited. There is generally a trade-off between missed links and false links (or recall and precision, or sensitivity and specificity (93) and no point of zero linkage error anywhere on the spectrum of agreement. The range of analysis outputs produced from a range of link-acceptance thresholds is not guaranteed to encompass the true value of any target parameter.

Bias analysis in which estimates of the likely or plausible extents of linkage error are used to either make qualitative inferences about the strength and direction of linkage error bias, or to quantitatively adjust analysis outputs for its influence linkage error bias (56,75,94). If information about how linkage error affects the distribution of outcomes and exposures is available, it may be possible to use well-established techniques for quantitative bias analysis, to adjust for these errors (88,95). The strengths of bias analysis lie in its flexibility; if empirical estimates of linkage error rates are unavailable then assumptions about these can be specified (96). Quantitative bias adjustment is particularly relevant for simple analyses but becomes more complex with complicated designs involving more than two data sources and/or a number of covariates (24).

2.7. VALIDATING LINKAGE RESULTS

The final step of the linkage process is the validation of the linked data. A number of validation routines can be applied to avoid incorrect data linkage and to ensure the high quality of the final dataset. Linked data validation methods include both plausibility checks within the primary data and consistency checks of information given in primary and secondary data (97). Another approach is to assess the extent to which the matched sample reflects the target population. For instance, in a study linking a single State's cancer registry to Medicare administrative claims for that State, researchers may use estimates of the percentage of cancer patients aged 65 years and older to determine what percentage of patients in the cancer registry would be expected to be linked to the Medicare data. If estimates indicate that 60 percent of cancer patients in the State are 65 years and older, then it is reasonable to expect that 60 percent of the patients in the cancer registry will be matched with Medicare. If, instead, the researcher finds that only 30 percent of patients in the cancer registry are successfully matched, this may serve as a signal that there is a problem with the matching algorithm (20). While not well-documented in the literature, some form of manual review is typically employed to check the results. Before starting the manual review process, a set of decision rules is developed to standardise the decision process across reviewers. Next, a random sample is drawn from the set of all potential matches identified during the blocking phase. Following the decision rules, one or more reviewers then determine whether each potential match is a match or non-

match. Finally, the decisions documented during the manual review process are used as a gold standard against which the decisions made by the algorithm are compared, allowing for the calculation of the sensitivity, PPV, and f-measure of the algorithm. A good algorithm should have scores of 95 percent or better across the three metrics (20). Other quality and completeness methods including comparisons with published reports have also been used in previous study to validate linked data (98).

2.8. CONCLUSIONS

The content and quality of the data sources to be linked play an important role in the choice of linkage methods. While deterministic methods are simplest and best suited to 'perfect' data where there are unique personal identifiers or highly discriminating linkage keys, probabilistic methods are more complex and can be adapted to imperfect data. For confidentiality and other reasons, it is generally considered good practice to separate the processes of data linkage (via trusted third parties) from the analysis of the linked data. However, the 'black box' of data linkage can make it difficult for researchers to judge the reliability of the resulting linked data for their intended purposes. Linkage errors, where records cannot be linked or are incorrectly linked, pose the greatest threat to the quality of the linked data and ultimately lead to information bias and selection bias. Care must be taken to assess the quality of the linkage in order to provide reliable results.

Researchers should be proactive in assessing linkage quality with respect to bias due to linkage errors, in understanding the consequences of underlying data quality and linkage errors, and in appropriately accounting for these in study design and analysis. Methods for handling linkage errors may lead to more robust research, but they are still an area of ongoing research. Finally, researchers should validate the linked data before undertaking any analysis using the data.

The next chapter of this thesis (chapter 3), in addition to describing the data sources used in the thesis, describes the practical implementation of HISlink using deterministic methods, the challenges and confidentiality issues encountered, the comparison of linked and unlinked data to assess the quality of HISlink data and identify potential sources of bias, and how linked data were validated using different methods, including checking for implausible scenarios, comparison with reference data (previous reports, previous linkages).

2.9. BIBLIOGRAPHY

1. Dunn HL. Record linkage. American Journal of Public Health and the Nation's Health. 1946;36(12):1412–6.
2. Boyd JH, Ferrante AM, O'Keefe CM, Bass AJ, Randall SM, Semmens JB. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. BMC Health Serv Res. 2012 Dec;12(1):480.
3. Centre for Health Record Linkage (CHeReL). New South Wales (NSW) Gouvernment Website - Centre for Health Record Linkage. [cited 2023 Feb 9]. How record linkage works. Available from: https://www.cherel.org.au/how-record-linkage-works#:~:text=How%20record%20linkage%20works,of%20health%20events%20for%20individuals.
4. Brook EL, Rosman DL, Holman CDJ. Public good through data linkage: measuring research outputs from the Western Australian Data Linkage System. Australian and New Zealand Journal of Public Health. 2008 Feb;32(1):19–23.
5. Swart E, Stallmann C, Powietzka J, March S. Datenlinkage von Primär- und Sekundärdaten: Ein Zugewinn auch für die kleinräumige Versorgungsforschung in Deutschland? Bundesgesundheitsbl. 2014 Feb;57(2):180–7.
6. Harron K, Mackay E, Elliot M. An introduction to data linkage. 2016;
7. Mason LG. A Comparison of Record Linkage Techniques November 2018.
8. Moravec H. When will computer hardware match the human brain?
9. Newcombe HB, Kennedy JM, Axford S, James AP. Automatic Linkage of Vital Records: Computers can be used to extract" follow-up" statistics of families from files of routine records. Science. 1959;130(3381):954–9.
10. Holman CDJ, Bass JA, Rosman DL, Smith MB, Semmens JB, Glasson EJ, et al. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. Aust Health Review. 2008;32(4):766.
11. Holman CDJ, Bass AJ, Rouse IL, Hobbs MST. Population-based linkage of health records in Western Australia: development of a health services research linked database. Australian and New Zealand Journal of Public Health. 1999 Oct;23(5):453–9.
12. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke JP, Ford DV, et al. The SAIL databank: linking multiple health and social care datasets. BMC Med Inform Decis Mak. 2009 Dec;9(1):3.
13. Furu K, Wettermark B, Andersen M, Martikainen JE, Almarsdottir AB, SÃ¸rensen HT. The Nordic Countries as a Cohort for Pharmacoepidemiological Research. Basic & Clinical Pharmacology & Toxicology. 2010 Feb;106(2):86–94.
14. Harron K. Data linkage in medical research. bmjmed. 2022 Mar;1(1):e000087.

15. Harron K, Gilbert R, Cromwell D, van der Meulen J. Linking Data for Mothers and Babies in De-Identified Electronic Health Data. Gebhardt S, editor. PLoS ONE. 2016 Oct 20;11(10): e0164667.
16. Maggi F. A Survey of Probabilistic Record Matching Models, Techniques and Tools.
17. Harron K, Goldstein H, Dibben C. Methodological developments in data linkage. John Wiley & Sons; 2015.
18. Christen P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer: Data-centric systems and applications.; 2012.
19. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future: Health Services Research and Data Linkages. Health Services Research. 2010 Oct;45(5p2):1468–88.
20. Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR. Linking data for health services research: a framework and instructional guide. 2014;
21. Boyd JH. Record Linkage Techniques: Exploring and developing data matching methods to create national record linkage infrastructure to support population level research. Curtin University; 2016.
22. Ranbaduge T. A Scalable Blocking Framework for Multidatabase Privacy-preserving Record Linkage. The Australian National University; 2018.
23. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. Big Data & Society. 2017 Dec;4(2):205395171774567.
24. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. International Journal of Epidemiology. 2017 Oct 1;46(5):1699–710.
25. Calderwood L, Lessof C. Enhancing Longitudinal Surveys by Linking to Administrative Data. In: Lynn P, editor. Methodology of Longitudinal Surveys [Internet]. Chichester, UK: John Wiley & Sons, Ltd; 2009 [cited 2020 Sep 29]. p. 55–72. Available from: http://doi.wiley.com/10.1002/9780470743874.ch4
26. Cadarette SM, Wong L. An Introduction to Health Care Administrative Data. CJHP [Internet]. 2015 Jun 25 [cited 2023 May 26];68(3). Available from: http://www.cjhp-online.ca/index.php/cjhp/article/view/1457
27. Kroeze H. Methodology: data_linkage [Internet]. 2017 [cited 2023 May 26]. Available from: https://cros-legacy.ec.europa.eu/system/files/s-dwh-m_4.2_methodology_data_linkage_v2.pdf
28. Briffa T, Heyworth J. Introductory analysis of linked health data course. Principles and hands-on applications. Version 3.5. 2019.
29. Doidge J, Christen P, Harron K. Quality assessment in data linkage. Joined up data in government: the future of data linking methods. 2020;

30. Kvalsvig A, Gibb S, Teng A. Linkage error and linkage bias: A guide for IDI users. University of Otago. 2019;
31. Jamieson E, Roberts J, Browne G. The feasibility and accuracy of anonymized record linkage to estimate shared clientele among three health and social service agencies. Methods of information in medicine. 1995;34(04):371–7.
32. Berkeley M. TEXTBOOK OF MEDICAL RECORD LINKAGE. The Journal of the Royal College of General Practitioners. 1987;37(304):518.
33. Ali MS, Ichihara MY, Lopes LC, Barbosa GCG, Pita R, Carreiro RP, et al. Administrative Data Linkage in Brazil: Potentials for Health Technology Assessment. Front Pharmacol. 2019 Sep 23;10:984.
34. Mears GD, Rosamond WD, Lohmeier C, Murphy C, O'Brien E, Asimos AW, et al. A Link to Improve Stroke Patient Care: A Successful Linkage Between a Statewide Emergency Medical Services Data System and a Stroke Registry: STROKE LINKAGE. Academic Emergency Medicine. 2010 Dec;17(12):1398–404.
35. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. Journal of Clinical Epidemiology. 2011 May;64(5):565–72.
36. Poluzzi E, Piccinni C, Carta P, Puccini A, Lanzoni M, Motola D, et al. Cardiovascular events in statin recipients: impact of adherence to treatment in a 3-year record linkage study. Eur J Clin Pharmacol. 2011 Apr;67(4):407–14.
37. Fleming M, Kirby B, Penny KI. Record linkage in Scotland and its applications to health research: *Record linkage in Scotland*. Journal of Clinical Nursing. 2012 Oct;21(19pt20):2711–21.
38. Van der Heyden J, Van Oyen H, Berger N, De Bacquer D, Van Herck K. Activity limitations predict health care expenditures in the general population in Belgium. BMC Public Health. 2015 Dec;15(1):267.
39. Van der Heyden J, Charafeddine R, De Bacquer D, Tafforeau J, Van Herck K. Regional differences in the validity of self-reported use of health care in Belgium: selection versus reporting bias. BMC Med Res Methodol. 2016 Dec;16(1):98.
40. Charafeddine R, Berger N, Demarest S, Van Oyen H. Using mortality follow-up of surveys to estimate social inequalities in healthy life years. Popul Health Metrics. 2014 Dec;12(1):13.
41. Health and Social Care Information Centre. Methodology for creation of the HES Patient ID (HESID) [Internet]. 2014 [cited 2023 May 26]. Available from: https://www.cl.cam.ac.uk/~rja14/Papers/HESID_Replacement_Nov09.pdf
42. Hagger-Johnson G, Harron K, Fleming T, Gilbert R, Goldstein H, Landy R, et al. Data linkage errors in hospital administrative data when applying a pseudonymisation algorithm to paediatric intensive care records. BMJ Open. 2015 Aug;5(8):e008118.

43. Grannis SJ, Overhage JM, McDonald CJ. Analysis of Identifier Performance using a Deterministic Linkage Algorithm.
44. Gomatam S, Carter R, Ariet M, Mitchell G. An empirical comparison of record linkage procedures. Statistics in medicine. 2002;21(10):1485–96.
45. Fellegi IP, Sunter AB. A theory for record linkage. Journal of the American Statistical Association. 1969;64(328):1183–210.
46. Joshua T, Stefan B. Record Linkage. In: Big Data and Social Science. Chapman and Hall/CRC; 2020. p. 43–65.
47. Winkler WE. Improved decision rules in the fellegi-sunter model of record linkage. Vol. 56. Bureau of the Census Washington, DC; 1993.
48. Jaro MA. Probabilistic linkage of large public health data files. Statistics in medicine. 1995;14(5-7):491–8.
49. Winkler WE. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. 1990 [cited 2023 May 26]; Available from: file:///C:/Users/FiBe649/Downloads/WinklerStringComparator1990_056%20(2).pdf
50. Wajda A, Roos LL. Simplifying record linkage: software and strategy. Computers in Biology and Medicine. 1987;17(4):239–48.
51. Winkler WE. Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. US Bureau of the Census Washington, DC; 2000.
52. Dey D, Sarkar S, De P. Entity matching in heterogeneous databases: a distance-based decision model. In: Proceedings of the Thirty-First Hawaii International Conference on System Sciences [Internet]. Kohala Coast, HI, USA: IEEE Comput. Soc; 1998 [cited 2023 May 26]. p. 305–13. Available from: http://ieeexplore.ieee.org/document/649225/
53. Foster I, Ghani R, Jarmin RS, Kreuter F, Lane J. Big Data and social science: Data science methods and tools for research and practice. CRC Press; 2020.
54. Bhattacharya I, Getoor L. Collective entity resolution in relational data. ACM Trans Knowl Discov Data. 2007 Mar;1(1):5.
55. Moore CL, Amin J, Gidding HF, Law MG. A New Method for Assessing How Sensitivity and Specificity of Linkage Studies Affects Estimation. Fernandez-Reyes D, editor. PLoS ONE. 2014 Jul 28;9(7):e103690.
56. Winglee M, Valliant R, Scheuren F. A case study in record linkage. Surv Methodol. 2005;31(1):3–11.
57. Nanayakkara C, Christen P, Ranbaduge T, Garrett E. Evaluation measure for group-based record linkage. IJPDS [Internet]. 2020 Oct 19 [cited 2023 May 26];4(1). Available from: https://ijpds.org/article/view/1127
58. Elfeky MG, Verykios VS, Elmagarmid AK. TAILOR: a record linkage toolbox. In: Proceedings 18th International Conference on Data Engineering [Internet]. San

Jose, CA, USA: IEEE Comput. Soc; 2002 [cited 2023 May 26]. p. 17–28. Available from: http://ieeexplore.ieee.org/document/994694/

59. Harron K, Doidge J. Challenges and opportunities in using administrative data linkage for research: the importance of quality assessment for understanding bias [Internet]. 2020 Jan; UCL Great Ormond Street Institute of Child Health. Available from: https://www.ucl.ac.uk/population-health-sciences/sites/population_health_sciences/files/1-nash-mina_katieharron_jan2020.pdf
60. Schnell R. Linking Surveys and Administrative Data [Internet]. 2013 [cited 2023 May 28]. Available from: file:///C:/Users/FiBe649/Downloads/SSRN-id3549220%20(3).pdf
61. Maastricht University, Netherlands, Künn S. The challenges of linking survey and administrative data. izawol [Internet]. 2015 [cited 2022 Mar 3]; Available from: http://wol.iza.org/articles/challenges-of-linking-survey-and-administrative-data
62. Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang LC, Smith P, et al. GUILD: GUidance for Information about Linking Data sets†. Journal of Public Health. 2018 Mar 1;40(1):191–8.
63. Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. Annals of Human Biology. 2020 Feb 17;47(2):218–26.
64. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, et al. Data Linkage: A powerful research tool with potential problems. BMC Health Serv Res. 2010 Dec;10(1):346.
65. Phillips M, Knoppers BM, Baker D. Privacy-Preserving Record Linkage: Ethico-Legal Considerations [Internet]. 2018 [cited 2023 May 28]. Available from: https://www.ga4gh.org/wp-content/uploads/2018-03-20-%E2%80%94-PPRL-legal-primer.pdf
66. Jutte DP, Roos LL, Brownell MD. Administrative Record Linkage as a Tool for Public Health Research. Annu Rev Public Health. 2011 Apr 21;32(1):91–108.
67. Green E, Ritchie F, Mytton J, Webber DJ, Deave T, Montgomery A, et al. Enabling data linkage to maximise the value of public health research data: Summary report. 2015;
68. European Parliament and European Council. Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) [Internet]. 2022 [cited 2022 Sep 27]. Available from: https://data.consilium.europa.eu/doc/document/PE-85-2021-INIT/en/pdf
69. Sakshaug JW, Couper MP, Ofstedal MB, Weir DR. Linking Survey and Administrative Records: Mechanisms of Consent. Sociological Methods & Research. 2012 Nov;41(4):535–69.

70. Sakshaug JW, Schmucker A, Kreuter F, Couper MP, Holtmann L. Respondent understanding of data linkage consent. Survey Methods: Insights from the Field (SMIF). 2021;
71. March S, Andrich S, Drepper J, Horenkamp-Sonntag D, Icks A, Ihle P, et al. Good Practice Data Linkage (GPD): A Translation of the German Version. IJERPH. 2020 Oct 27;17(21):7852.
72. March S. Individual Data Linkage of Survey Data with Claims Data in Germany—An Overview Based on a Cohort Study. IJERPH. 2017 Dec 9;14(12):1543.
73. van Veen EB. Observational health research in Europe: understanding the General Data Protection Regulation and underlying debate. European Journal of Cancer. 2018 Nov;104:70–80.
74. Kelman CW, Bass AJ, Holman CDJ. Research use of linked health data — a best practice protocol. Australian and New Zealand Journal of Public Health. 2002 Jun;26(3):251–5.
75. Doidge JC, Harron KL. Reflections on modern methods: linkage error bias. International Journal of Epidemiology. 2019 Oct 21;dyz203.
76. Sariyar M, Borg A, Pommerening K. Missing values in deduplication of electronic patient data. Journal of the American Medical Informatics Association. 2012 Jun 1;19(e1):e76–82.
77. Baldi I, Ponti A, Zanetti R, Ciccone G, Merletti F, Gregori D. The impact of record-linkage bias in the Cox model. Journal of Evaluation in Clinical Practice. 2010 Feb;16(1):92–6.
78. Neter J, Maynes ES, Ramanathan R. The Effect of Mismatching on the Measurement of Response Errors. Journal of the American Statistical Association. 1965 Dec;60(312):1005–27.
79. Ford JB, Roberts CL, Taylor LK. Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data. Paediatr Perinat Epidemiol. 2006 Jul;20(4):329–37.
80. Bohensky M. Bias in data linkage studies. Methodological developments in data linkage. 2015;63–82.
81. Lariscy J. Differential Record Linkage by Hispanic Ethnicity and Age in Linked Mortality Studies: Implications for the Epidemiologic Paradox [Internet]. SocArXiv; 2019 Dec [cited 2023 May 29]. Available from: https://osf.io/tw9a4
82. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. International Journal of Epidemiology. 2010 Apr;39(2):417–20.
83. VanderWeele TJ. Commentary: Resolutions of the birthweight paradox: competing explanations and analytical insights. International Journal of Epidemiology. 2014 Oct;43(5):1368–73.

84. Rentsch CT, Harron K, Urassa M, Todd J, Reniers G, Zaba B. Impact of linkage quality on inferences drawn from analyses using data with high rates of linkage errors in rural Tanzania. BMC Med Res Methodol. 2018 Dec;18(1):165.

85. Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. BMC Med Res Methodol. 2014 Dec;14(1):36.

86. Silveira DP da, Artmann E. Accuracy of probabilistic record linkage applied to health databases: systematic review. Revista de saúde pública. 2009;43:875–82.

87. Christen P, Goiser K. Assessing deduplication and data linkage quality: what to measure? In: 4 th Australasian Data Mining Conference [Internet]. Sydney, Australia; 2005 [cited 2023 May 29]. p. 11. Available from: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/32722.pdf#page=45

88. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. International Journal of Epidemiology. 2014 Dec 1;43(6):1969–85.

89. Goldstein H, Harron K, Wade A. The analysis of record-linked data using multiple imputation with data value priors. Statist Med. 2012 Dec 10;31(28):3481–93.

90. Fawcett J, Blakely T, Atkinson J. Weighting the 81, 86, 91 & 96 census-mortality cohorts to adjust for linkage bias. Department of Public Health, Wellington School of Medicine and Health …; 2002.

91. Chipperfield J. A weighting approach to making inference with probabilistically linked data. Statistica Neerlandica. 2019;73(3):333–50.

92. Chipperfield J. Bootstrap inference using estimating equations and data that are linked with complex probabilistic algorithms. Statistica Neerlandica. 2020;74(2):96–111.

93. Smalheiser NR, Torvik VI. Author name disambiguation. Ann Rev Info Sci Tech. 2009;43(1):1–43.

94. Lahiri P, Larsen MD. Regression analysis with linked data. Journal of the American statistical association. 2005;100(469):222–30.

95. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidem Drug Safe. 2006 May;15(5):291–303.

96. Doidge JC, Morris JK, Harron KL, Stevens S, Gilbert R. Prevalence of Down's Syndrome in England, 1998–2013: Comparison of linked surveillance data and electronic health records. IJPDS [Internet]. 2020 Mar 19 [cited 2023 May 29];5(1). Available from: https://ijpds.org/article/view/1157

97. Druschke D, Arnold K, Heinrich L, Reichert J, Rüdiger M, Schmitt J. Individual-Level Linkage of Primary and Secondary Data from Three Sources for

Comprehensive Analyses of Low Birthweight Effects. Gesundheitswesen. 2020 Mar;82(S 02):S108–16.

98. Kidd JBR. Developing a population data linkage cohort to investigate the impact on child oral health outcomes following the roll-out of the Childsmile programme in Scotland [Internet]. [Scotland]: University of Glasgow; 2019 [cited 2023 May 29]. Available from: file:///C:/Users/FiBe649/Desktop/Jamie%20Brian%20Rutherford%20Kidd_2019.pdf

CHAPTER 3. DATA SOURCES AND IMPLEMENTATION OF THE LINKAGE

3.1. DATA SOURCES

HISlink involves the linkage of two data sources: the BHIS data and the BCHI. A first linkage with BHIS 2008 data was done in 2012. Based on the experience of this project a more systematic linkage between BHIS and BCHI data was set up, starting from the BHIS 2013 onwards. HISlink specifically refers to the latter and studies presented in this thesis are based on the linkages carried out in the framework of HISlink.

3.1.1. Belgian Health interview survey

The BHIS was first launched in 1997 and since then it has been organised with intervals between three and five years. More specifically, the following waves have been organised: BHIS 1997, 2001, 2004, 2008, 2013, 2018 and the BHIS 2023 is currently underway. A brief overview of the most essential features of the BHIS are provided here. Detailed information on the BHIS can be found in the survey protocols available on the BHIS website (1) and in Demarest et al. (2013) (2).

3.1.1.1 Organisational and legal context

The BHIS is executed by Sciensano, the Belgian health institute, and it has been commissioned and co-financed by the different Belgian authorities competent in the field of public health in the framework of interministerial agreements between the Belgian Federal State and the Federated authorities (Regions and Communities). Important partners are Statbel, the Belgian statistical office which is playing an essential role in the field work and data collection, and the Center for Statistics (Censtat) of the UHasselt, for statistical advice. There is also a scientific steering committee with representatives from universities, administrations and other stakeholders.

3.1.1.2 Aims

The main objective of the BHIS is to measure the health status of the population in Belgium, accounting also for three sub-regional populations (Flemish, Walloon and Brussels). The BHIS is designed to obtain information on people's health experience, their attitudes and health-related behaviours, the extent to which they use healthcare facilities and their use of preventive health and social services. This information

enables health authorities and stakeholders to pursue a proactive health policy aimed at improving public health and addressing the needs of groups at risk, but also to influence policy and health programmes with surveillance data. The BHIS provides an overall picture of the health status of the population and allows identifying of the main health problems, as well as the social and behavioural factors that influence them. The repeated organisation of the BHIS enables the studying of trends in public health-related indicators and contributing to policy evaluation. Through the BHIS, a rich database is constructed for the scientific community allowing more in-depth research. The information collected via the BHIS is not only used at regional and national level but also for reporting to international instances such as Eurostat, World Health Organization (WHO), United Nations (UN) and the Organisation for Economic Co-operation and Development (OECD).

3.1.1.3 Target population, sample size and sample method

The target population of the BHIS includes all persons residing in Belgium, regardless of their age, place of birth, nationality or other characteristics. The sampling frame consists of all households listed in the National Register (NR). Specifically, a household is defined as the people living at the address of a reference person. Collective households are included in the sampling frame, with each individual belonging to a collective household being considered as a one-person household. However, people living in:

- an institution (including psychiatric institutions), with the exception of elderly people living in nursing homes,
- a religious community or cloister with more than 8 persons,
- a prison

are post hoc excluded from the survey for practical reasons.

The basic net-sample size of BHIS, expressed as the number of successful interviews to be obtained is defined before starting data collection, taking into account specific budget constraints and the available logistic means. For all BHIS up to 2013, the total number of successful interviews for the basic sample was set at 10,000 (3,500 for Flanders, 3,500 for Wallonia, including 300 for East Belgium and 3,000 for Brussels Capital Region). From the BHIS 2018 and onwards, the basic net-sample size

increased to 10,700 since the net-sample size for the Flemish Region increased to 4,200 successful interviews. This increase was requested to obtain some minimal information at the level of the health region "*zorgregio's*" (health regions include different municipalities and can be interpreted as a level between province and municipality). From the Protocol Agreement BHIS 2001 onwards, other authorities than the Commissioners could ask (and pay) for (a) supplement sample(s) as long as the total size of this (these) supplement sample(s) did not exceed 3,000 participants, this in order to keep the fieldwork manageable. Oversampling of specific population groups was conducted for specific provinces in 2001, 2004, 2013, 2018 and 2023. Table 3.1 gives an overview of the oversampling for all survey editions and total interviews (planned and realised).

Table 3.1: Overview of the sample size of the Belgian health interview surveys 1997-2023

Year	**1997**	**2001**	**2004**	**2008**	**2013**	**2018**	**2023**
Basic sample	10000	10000	10000	10000	10000	10700	10700
Provincial oversampling							
Antwerp		350					
Hainaut		500					
Limburg		200	450				
Luxembourg		1000	897		600		
Oversampling German Community						600	600
Oversampling elderly							
65-84 years			550				
75-84 years				400			
85 years +			700	850			
Total interviews							
Planned	10000	12050	12597	11250	10600	11300	11300
Realized	10221	12111	12945	11254	10829	11611	N.A*

**NA = Not available. As data collection for the BHIS 2023 is currently underway, the final sample size, i.e. the number of interviews conducted, is not yet available.*

In order to achieve the predefined number of successful interviews, taking into account all the technical constraints relative to the data collection mode as well as financial consideration, a multistage sampling design is developed. In summary, the

final sampling scheme, i.e. the mechanism to obtain a probabilistic sample of households and respondents, is a combination of several sampling techniques: stratification, multistage sampling and clustering. The selection process consists of the following steps:

1. Regional stratification where the number of interviews to be carried out for each region is fixed at 4,200 for Flanders (since BHIS 2018, 3,500 in the previous editions), 3,500 for Wallonia and 3,000 for Brussels. These figures do not include the oversampling. The reason for this stratification is to ensure that inferences can be drew for each region with nearly the same precision.

2. Stratification at the level of the provinces. This second level of stratification is done to improve the quality of the sample over a simple random sample. In particular, a balanced geographical spread is achieved. For the base sample, the sample size within the provincial stratification is proportional to the population size of the province.

3. Stratification at the level of the zorgregio's/arrondissements électoraux. Since the BHIS 2018 an extra stratification level – the level of *zorgregio's* – was added to the sample scheme. This third level of stratification was introduced on the demand of the Flemish community who wished to make geographical comparisons at a lower geographical level (*zorgregio's*). To have a consistent methodology it was explored which geographical level could be identified in the Walloon region, having more or less the same number of geographical units. This appeared to be the *arrondissements électoraux*. Since neither *zorgregio's* nor *arrondissements électoraux* trespass provincial borders, this additional stratification level does not impact the provincial stratification. Ultimately, 29 strata were distinguished in the sampling procedure:

- 14 strata in the Flemish Region (the 14 '*zorgregio's*'),
- 1 stratum in the Brussels Capital Region,
- 14 strata in the Walloon region: the 14 '*arrondissements électoraux*' (except the 9 municipalities of the German Community),
- 1 stratum for the German Community.

4. Then, within the strata, units are accessed in two (for the households (HH)) or three (for the individuals) stages:

(1) Municipalities are selected with a selection probability proportional to their size, within each stratum. These municipalities are called the Primary Sampling Units (PSU). To facilitate the fieldwork, for each PSU selected, a group of 50 individuals residing in that municipality must be interviewed successfully during the survey year.

(2) Within each municipality, a sample of households - the Secondary Sampling Units (SSU) - is drawn in such a way that 50 individuals per PSU can be interviewed in total.

(3) Finally, at most four individuals - the Tertiary Sampling Units (TSU) - are chosen for the interviews within each household. Only questioning the reference person within a household would not enable us to give a good picture of a household's health status. For households with four members or less, all the members are selected. For households with at least five members, the reference person and his/her partner (if any) are selected. Among the remaining household members, a random selection is made, so as to yield four selected household members. Interviewing more than four persons within a household is inefficient because of the familial correlation and because the burden on the household would be too great.

5. To avoid seasonal effects, interviews are spread over the whole calendar year so that each quarter is comparable in terms of number of selected units. The quarters are defined as follows: Q1: January-March; Q2: April-June; Q3: July-September and Q4: October-December.

3.1.1.4 Field work and data collection

The fieldwork is organised by Sciensano in collaboration with Statbel, the Belgian Statistical Office. The selected households for the BHIS are informed by means of an introduction letter which contains information about the commissioners, the aims of the BHIS and the voluntary character of the BHIS. Furthermore, this letter states that an interviewer from Statbel will contact them. This introduction letter is accompanied by a leaflet which contains more detailed information on the BHIS. With the exception

of the BHIS 2013 wave, no incentives for participation are foreseen. Interviews are carried out by around 200 trained interviewers at the respondent's home. If a household cannot be contacted after making at least five contact attempts or refuses to participate, the interviewer receives a replacement or substitute household from Statbel. The first three substitute households are similar to the first one in terms of age of the reference person, statistical sector of residence and household size. If the fourth household also does not participate, there is an extra reserve of four additional households, but these do not necessarily have the same characteristics as the initially selected one. Monitoring and follow-up is done by a central secretariat at Statbel.

Prior to contact by the interviewer, each selected household receives a letter and information leaflet. Questionnaires are available in the three national languages (Dutch, French, German) and in English. The questions in the questionnaires are organised in terms of modules, i.e. a set of questions related to specific topics. The interview consists of several parts that make use of different modes of data collection. Household information (i.e. composition of the household, household income, expenses on healthcare consumption, characteristics of the house, etc.) is obtained from the reference person or the partner through a face-to-face (F2F) interview. An F2F interview is also conducted with each selected member of the household (max. 4 household members are interviewed) to obtain information on health perception, chronic conditions, healthcare consumption, etc. If the selected person is younger than 15 years old or not able to answer him/herself, a proxy interview is conducted. Questions that are more sensitive are addressed through self-administered paper and pencil (P&P) questionnaire, which is restricted to people aged 15 years and older. This self-administered P&P questionnaire contains questions on mental health, alcohol consumption, illicit drug use, etc. The use of this second and more private mode containing a specific part of the questionnaire is a type of mixed-mode design used to reduce the social desirability bias for sensitive questions and so the overall measurement error (3,4). The P&P is not completed in the case of a proxy interview. Until 2008, the F2F interview was conducted via a Paper Assisted Personal Interviewing (PAPI). In 2013, the switch was made to a Computer Assisted Personal Interviewing (CAPI). Data obtained via CAPI are transferred to a server of our fieldwork partner Statbel through a secured Internet connection and the P&P questionnaires are sent (through postal mail) or brought to Statbel for encoding. From

BHIS 2018, a module containing questions on the mental health of children and adolescents was introduced in the BHIS. The questions had to be completed via a Computer Assisted Personal Interview (CASI). This means that respondents entered their own answers on the laptop of the interviewer. This sub-module had to be completed by the parents of children from 2 to 15 years or by the adolescents between 15 and 18 years themselves.

3.1.1.5 Strengths and limitations of BHIS data

From previous sections, there is no doubt that the BHIS is an important source of information on population health in Belgium. It should of course also be acknowledged that there are limitations as in all the HIS. Table 3.2 presents an overview of the main strengths and limitations of the (B)HIS data in general.

Table 3.2: Overview of the strengths and limitations of the (Belgian) health interview survey data

Strengths	Limitations
• Data are collected at the level of the total population, including people who do not make use of health services.	• Expensive (but it is relatively cheap compared with other surveys such as health examination survey)
• Information is obtained from the perspective of the individual him/herself.	• Subject to biases such as selection bias, recall bias or social desirability bias
• The collection of self-perceived health, lifestyle, behaviour is only (or mainly) possible through a survey.	• Logistically more demanding and time-consuming compared with administrative data
• Information is collected simultaneously on the health status, health behaviour and healthcare utilisation of individuals, but also on socio-demographic health determinants, such as e.g. socioeconomic status.	• A limited number of questions can be included because the burden for interviewers and interviewees must remain acceptable. Otherwise, this might yield a lower participation rate
• This horizontal data collection makes it possible to study the relation between different domains and topics.	• Representativity is a concern in the case of low response rates.

3.1.2. Belgian Compulsory Health insurance

In Belgium, there is compulsory health insurance with exhaustive and detailed data on the reimbursed health expenses of over 99% of the total population. This means

that almost every citizen holds a membership at one of the seven sickness funds. However, there are some differences in coverage rates between regions and demographic characteristics (5). Since 2002, the IMA, an overarching national organisation, collects and manages data from these sickness funds for all Belgian citizens (hereinafter referred to as BCHI data). The BCHI database is a longitudinal linkage at individual level with the following information:

1) Some socio-demographic information such as age, sex, place of residence, vital status (deceased yes/no and date of death), limited socioeconomic information, including the individual's status with respect to the entitlement to some social benefits, preferential reimbursement, etc.
2) Detailed information on reimbursed health care.
3) Detailed information on the delivery of reimbursed medicines (Pharmanet data).

The different components are linked together using a TTP (6), i.e. the linkage was outsourced to another organisation that has access to identifiable data and has performed the linkage. The database includes an arbitrary id-code, allocated by the TTP and is updated annually. Detailed information on the BCHI data can be found in (7).

As the primary goal of the BCHI data is for reimbursement purposes, the data on healthcare utilisation is highly accurate. BCHI data are widely used by important actors in the health field, such as the NIHDI, the KCE, the Belgian Federal Planning Bureau and the healthcare insurers for reimbursement purposes, assessment and planning of healthcare costs. In addition, BCHI data are also used for specific studies beyond their initial intended use (secondary use). An advantage is that the data are not self-reported or limited to a certain registration period but are continuously collected for administrative use. Although BCHI data do not include information on the diagnosis, methods have been developed to use those data to estimate the prevalence of certain chronic diseases at the level of the general population (pseudopathologies derived from medication use) (8). Due to a number of limitations to pseudopathologies / disease groups (uncertainty about the difference between pseudopathology / disease groups, lack of updates to outdated definitions, hospital drugs not taken into account when determining disease groups), the concept of

pseudopathologies has recently been revised and replaced by the term “Pharmacy Cost Groups” (PCG). The term “Pharmacy Cost Groups” is based on the Dutch pharmaceutical cost groups classification managed by the National Institute of Health Care and covers the burden better than disease groups (9). Furthermore, since BCHI data registrations are usually standardised and continuously collected, they enable trend analyses and longitudinal studies (10,11). However, BCHI data have some shortcomings: it only includes information on covered health services and goods, and there is limited information on outpatient supplements. In addition, there is limited socio-demographic data. Moreover, as BCHI data is based on billing of services which may involve several manipulations, data may be subject to errors (e.g. inaccurate procedure codes, upcoding errors, duplicate billing). Table 3.3 summarises the main strengths and limitations of BCHI data.

Table 3.3: Overview of the strengths and limitations of the Belgian Compulsory Health Insurance data

Strengths	Limitations
• Objective health consumption data	• Does not include information on the diagnosis
• For the whole population, no selection bias	• Only data regarding reimbursed healthcare consumption are considered
• Standardised and continuously collected	• Limited socio-demographic information
	• May be subject to errors (e.g. recording errors), missing data

Since 2002 a legal framework exists to use a permanent sample of the BCHI data for policy and research purposes. This database, officially called the “*Echantillon Permanent/Permanente Steekproef*” (EPS), is a representative randomised sample of 1/40th of the Belgian population. For the population of 65 years and older an extra sample is taken, as a result of which 1/20th of the population is included.

All the studies of this thesis were based on an individual linkage between the BHIS 2013 and BHIS 2018 with BCHI data using the national register number. In one of the studies not only linked data were used, but also results from the EPS to assess the selection bias of mammography uptake among women aged 50-69 years old.

3.2. IMPLEMENTATION OF THE LINKAGE

3.2.1. Context, commissioner and objectives

Health data are essential for the development of a coherent health policy. In Belgium, a lot of such data are available, including the BHIS and the BCHI data sources. A problem is that these data are often fragmented and not integrated into an effective integrated health information system. The BHIS and the BCHI data are two complementary cornerstones of the Belgian health information system. Self-reported information is collected during the BHIS on the health status, lifestyle, healthcare use and socio-demographic background characteristics of a representative sample of the population in Belgium. The BCHI database is an administrative database with detailed information on reimbursed healthcare expenses of the total population. As highlighted above, both data sources have their strengths and shortcomings. The HISlink project tries to overcome these shortcomings. Through an individual linkage of BHIS and BCHI data, it is possible to address health-policy-relevant questions that each of the data sources separately cannot answer. The linked database has the advantage of combining the strengths of both sources (horizontal data collection from the BHIS and detailed data on healthcare expenditure from BCHI data), which makes it a very powerful instrument for answering a number of policy-relevant questions.

In 2012, a series of NIHDI's reports had shown that healthcare expenditure was systematically lower in the Brussels capital region as compared with the Flemish and Walloon regions. An in-depth analysis of this phenomenon required more comprehensive data. However, BCHI data lack sociodemographic and health information which could be useful to understand the differences between the three regions. Sciensano was therefore commissioned by NIDHI in 2013 to explore this phenomenon in more detail by means of a linkage between the BHIS and BCHI data to verify whether these differences are related to the specific demographic, socio-economic and health characteristics of the inhabitants of these regions. In addition, the NIHDI expressed the wish to further explore the concerns of access to health care by the mean of the linked data. Indeed, the results of the BHIS 2008 had shown a sharp increase between 2004 and 2008 in the number of people indicating that they had had to postpone their health expenditure for financial reasons. Moreover, the NIHDI also requested to estimate the cost for the Belgian health insurance if some

groups of non-reimbursed medicines (analgesics, laxatives and calcium supplements) were to be reimbursed.

Next to this, BHIS is facing significant challenges due to the increasing demand of the commissioners to inflate the content of the questionnaire. Moreover, researchers need more detailed and complete data to perform more accurate analyses in order to draw valid conclusions. However, the perpetual search for more complete, comprehensive and complex data often leads to longer questionnaires, more tedious interviews and, as a consequence, an increasing burden for interviewers and interviewees, resulting also in decreased quality of the data. The HISlink is an effective way to substitute and supplement BHIS information with BCHI information, specifically on the use of reimbursed health care and medicines, in order to get more comprehensive and high-quality data without increasing the workloads for interviewers and for interviewees and to reduce the cost burden of obtaining additional information, given the expense of active follow-up of survey respondents.

In the past decade some projects have been carried out in which BHIS data have also been linked with mortality (12,13) data and census data (14). Those linked databases have proved to be powerful instruments to answer specific health research questions. They also enabled the acquiring of important know-how on the technical organisation, legal framework and privacy issues that are related with such a linkage. Up to now those linkages have been done in the framework of specific projects. No reflection has been made on a systematic linkage of BHIS data with administrative health data as a standard procedure of the organisation of a BHIS. However, such a systematic linkage could provide health policymakers with a powerful tool for public health research with direct relevance for health policy planning and evaluation.

The HISlink project is specifically meant to respond to policy-relevant questions raised by NIDHI (who is the commissioner of this project). Domains covered are, among others, socio-demographic differences in use of health care, evaluation of chronic morbidity indicators used by NIDHI and use of non-reimbursed medicines. More specifically, the linked data are used:

- to study access to health care
- to explore specific questions with respect to the use of medicines (including non-reimbursed medicines)

- for data evaluation, validation
- to study the determinants of health care utilisation
- to construct new IMA-based indicators in the framework of BHIS reports
- for population health monitoring, health surveillance and planning
- for healthcare research and quality of care assessment
- for other specific NIHDI research questions
- for studies assessing the economic impact of diseases and ill health.

These objectives are not exhaustive and may change over time depending on the specific objectives of the projects using the linked data.

3.2.2. History of HISlink

The BHIS and BCHI data are individually linked using the national register number (NRN). In Belgium, the NRN is a unique and personal identifier consisting of 11 digits used to identify each citizen who holds a Belgian identity document or a Belgian residence document. The NRN allows access to almost all “administrative” services including those related to healthcare service use.

Furthermore, Belgium has a central databank, the National population register (NR) which includes Belgians residing in Belgium and residing abroad if they are registered in diplomatic posts, but also non-Belgians, who officially reside in Belgium, and non-Belgians who declare to be refugees or are officially recognised as refugees. Data are collected at the level of the municipality and sent to the central registration office. Statbel receives regular updates of the complete population register. The core information in the NR includes given name(s) and surname, place and date of birth, sex, nationality, address of residence, place and date of death, civil status and household composition. Each individual in the NR is identified through the unique NRN. The use of information from the NR is strictly regulated by law. As there is a link to the NRN in many official databases, e.g. social security databases, mortality data, hospital discharge data, BCHI data, and as NR is the sample frame for the

selection of participants for various population surveys, including the BHIS, it is an ideal tool for the linkage of databases.

The first linkage was performed in 2013 with the BHIS 2008 data as a feasibility study (15). The current HISlink project was launched in 2017 on a systematic basis. So far, new linkages have been performed with data from the BHIS 2013 and the BHIS 2018. The linkage procedure with data from the BHIS 2023 is under preparation.

3.2.3. Partners involved

The HISlink project involves several partners among whom especially Statbel, IMA and eHealth have a key role.

- National Institute for Health and Disabilities Insurance (NIHDI)

The NIHDI, and more specifically the Health Care Service, Directorate for Research, Development and Quality promotion (RDQ), is the sponsor of the HISlink project.

- Sciensano

Sciensano, the Belgian institute for public health is responsible for carrying out the HISlink project, but also for the organisation of the BHIS on behalf of all Belgian authorities responsible for public health at federal, regional and community levels.

- Statbel

Statbel, the Belgian statistical office, acts as a subcontractor and is responsible for sampling and fieldwork for the BHIS under the instructions of Sciensano. For the HISlink, Statbel is responsible of the BHIS data encryption.

- National Intermutualist College

The National Intermutualist College (NIC) groups the sickness funds representing the entire Belgian socially insured population and hosting the healthcare consumption data. The NIC provides healthcare consumption data in this project.

- Healthcare insurers

In the framework of this project, the healthcare insurers or sickness funds provide eHealth with comprehensive information on all reimbursed medicines dispensed in public pharmacies and healthcare use data.

- InterMutualistic Agency (IMA)

IMA is a national overarching organisation that collects data from the seven healthcare insurers for all Belgian citizens. The Pharmanet data are provided by IMA, as well as the population data. IMA is also conducting a small cell risk analysis (SCRA). Once the data is encrypted by the TTP, IMA makes the data available to Sciensano.

- Crossroads Bank of Social Security

The Crossroads Bank of Social Security (CBSS) acts as TTP between IMA and the healthcare insurers in the framework of the present linkage project.

- eHealth

eHealth acts as a TTP and through the secure eHealthbox, the eHealth platform is used to exchange encrypted data between the partners. The eHealth platform also stores the files on a virtual hard disk with the Veracrypt software during the linking process.

3.2.4. Linkage process and data flow

The Figure 3.1 below depicts the data flow from HISlink 2018. The same flow was used in HISlink 2013. The linkage process is complex and requires several coding steps to ensure privacy and data protection. Overall, the process consists of two phases: the selection phase during which BHIS participants are selected and enriched with additional household members from IMA data warehouse (IMA DWH), and the data phase which consist of the actual linkage and involves data encryption and exchanges between health insurance funds, IMA and TTPs. In IMA DWH, BHIS participants are enriched with all other participating household members according to the MAF definition, irrespective of their individual participation in the survey. The MAF (i.e. “Maximum A Facturer” or Maximum billing) is a system introduced in 2002 that puts a ceiling on the total amount of co-payments (not supplements) at the level of a household during a calendar year, where the ceiling is a function of the household income. As the composition of households in the BHIS may be different from the composition of a MAF household, this enrichment is necessary to create relevant IMA indicators at household level. For example, the postponement of medical consumption in the BHIS is assessed at household level, so it is necessary to aggregate the IMA data (based on individual-level data, linked via the MAF household) to household level.

More specifically, the two phases can be summarised as follows:

In the selection phase:

1) Statbel selects the NRN of BHIS participants and transmits them to the Security Advisor of the NIC (step 1).
2) The NIC Security Advisor converts the NRN to C1 and transmits the list of C1 to the TTP CBSS (step 2).
3) The TTP CBSS converts the C1 into C2 and sends the list of C2 to the IMA (step 3). The composition of the household MAF is consulted in the IMA DWH. On the basis of this consultation, the selection list is enriched with an additional number of persons.
4) IMA transmits the enriched C2(2) list to the TTP CBSS (step 4).
5) The TTP CBSS converts the C2(2) list to C1(2) and sends the C1 list to the NIC Security Advisor (step 5).
6) The NIC Security Advisor converts C1(2) to NRN(2) and forwards the list of NRN to Statbel (step 6).

At the data phase:

1) Statbel transmits the enriched selection of NRN(2) to the TTP eHealth with an internal RN (Random Number) specific to this project (7.1). The NIC Security Advisor transmits an NRN/C1-encoded list of persons to the TTP eHealth, with C1 encrypted (7.2).
2) The TTP eHealth sends via the secure eHbox Cproject/RN to the TTP-CBSS (8.1). The TTP eHealth sends via the Cproject/C1 secure eHbox to the TTP CBSS (8.2). Statbel transmits BHIS data on an NR basis to the TTP CBSS (8.3).
3) On the basis of a second coding (C1 → C2), the data are selected from the IMA DWH (step 9).
4) The data is sent back on a C2 basis to the TTP CBSS (step 10).
5) TTP CBSS replaces C2 with Cproject and also converts the received data into Cproject. These are transmitted to the IMA DWH (step 11).
6) A small cell risk analysis (SCRA) is carried out by the IMA (step 12).

7) The data sets are made available to Sciensano researchers (Cproject) on IMA server step 13).

The pseudomised data are accessible through a Virtual Private Network (VPN) connection with secure token. Ultimately, a quadruple coding system ensured a coded database where no single party held all of the respective keys enabling identification of individual patients.

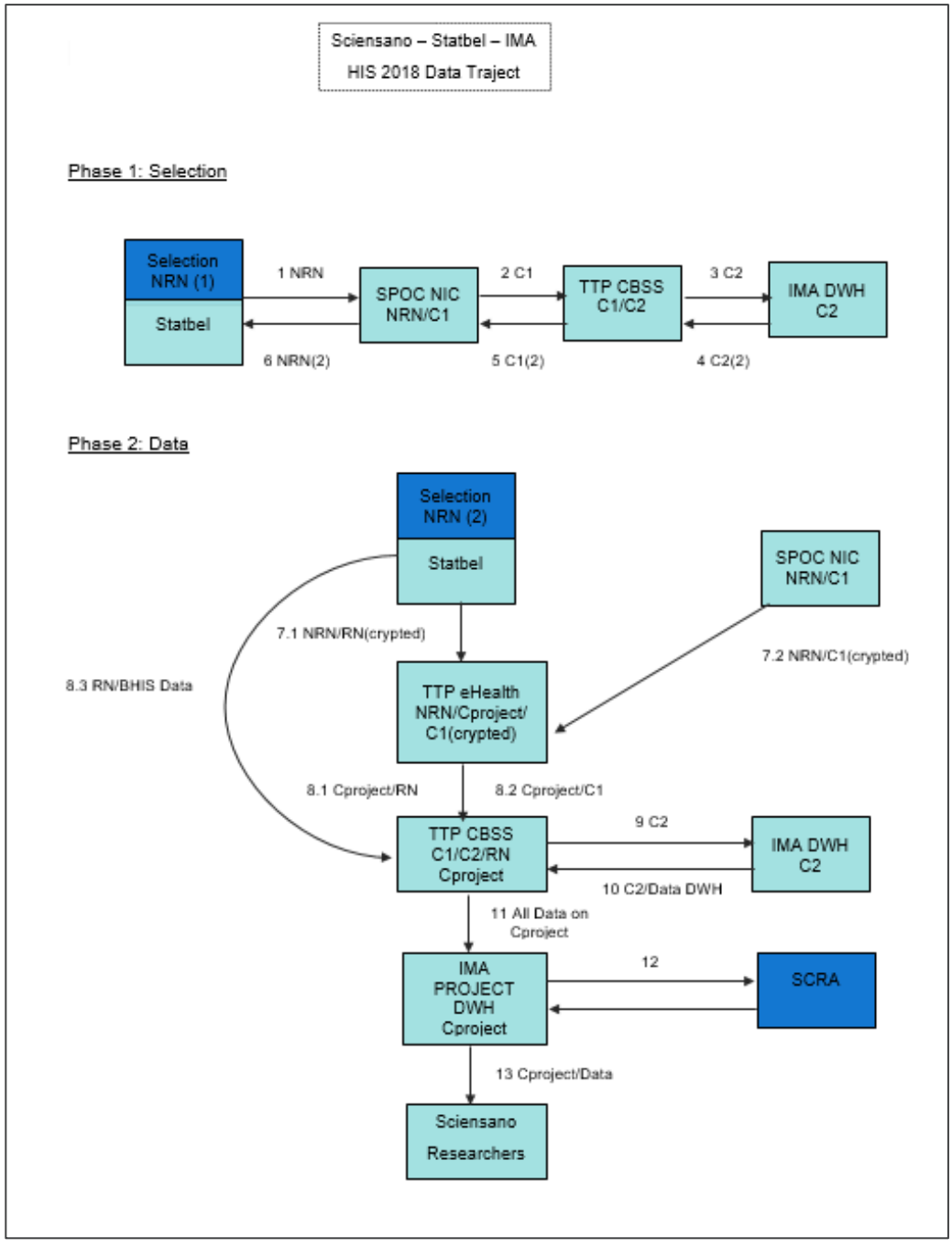


Figure 3.1: Step-by-step overview of linkage procedure and data coding system to enable data linkage for the HISlink 2018, Belgium

3.2.5. Ethics and privacy procedures

The BHIS 2013 and BHIS 2018 were carried out in line with the Belgian privacy legislation and were approved by the ethics committee of the University hospital of Ghent on October 1, 2012 (advice EC UZG 2012/658) and December 21, 2017 (advice EC UZG 2017/1454) respectively. The participation in the BHIS is voluntary. No written consent was foreseen. Participation was equivalent to giving consent. For the linkage with BCHI data, an authorisation was obtained from the Belgian Information security committee (ISC) acting as institutional review board (IRB) (local reference: Deliberation No. 17/119 of December 19, 2017, amended on September 3, 2019 for the HISlink 2013 and local reference: Deliberation No 20/204 of November 3, 2020 for the HISlink 2018). In its deliberation, the IRB required Sciensano to inform the BHIS participants about the linkage of their data. In view of the disproportionate effort to do so (almost 11,000 individuals for the BHIS 2013 and more than 12,000 individuals for the BHIS 2018), and because the linkage process was launched before the implementation of the General Data Protection Regulation (GDPR), Sciensano presented an alternative approach to IRB which was accepted. This approach consisted of an exemption from the obligation to provide information at the level of each participant, but a communication to the general public about the data processing through publication on the BHIS website. This communication mentioned the following elements: the name and address of the data controller, the precise purposes of the processing, the existence of a right of access and rectification of the data and the existence of a right of objection by the data subject, and the modalities for exercising these rights, the categories of data concerned, their origin and recipients.

3.2.6. Contents of the linked databases

The information collected in the BHIS includes data on health status, lifestyle and health behaviour, prevention and attitudes towards health, health consumption, social and environmental aspects of health and socio-demographic characteristics. The questions are organised by modules containing a set of questions related to the same topic. These topics are based on public health relevance and are selected in consultation with the commissioners. A core set of questions is repeated over time to

assess time trends. From the BCHI, objective information relative to healthcare consumption as well as a limited number of socio-demographic information is gathered. The linkage finally resulted in datasets containing for:

- HISlink 2018: about 1,232 variables and related indicators from BHIS and 133 variables from BCHI;
- HISlink 2013: 1,265 variables and related indicators from BHIS and 127 variables from BCHI.

Table A1 in annex presents an overview of the contents of the linked data.

3.2.7. Data flow and overall result of the linkage

All BHIS participants were eligible for inclusion in the HISlink. Figure 3.2 and Figure 3.3 present the selection process for final BHIS 2013 and BHIS 2018 participants, respectively.

HISlink 2013

In 2013, 5,055 households participated in the BHIS. From the 11,614 individuals belonging to those households, 10,829 actually participated in the survey and 785 were not invited for participation (as maximum 4 household members can participate). For the linkage Statbel managed to retrieve the NRNs of 11,226 individuals belonging to the participating households, including 1 duplicate. Those NRNs were sent through a TTP to IMA. At the IMA level, 10,699 records were retrieved and 527 were not. On the basis of the MAF household composition in the IMA DWH, 680 extra individuals were added leading to a total of 11,379 records. In total, the HISlink 2013 contained 12,294 records (including 1 duplicate) (see Figure 3.2). The overall linkage rate among individuals belonging to the participating households was 92.1% (10,699 out of 11,614). Among individuals who actually participated in the survey this percentage was 92.3% (9,998 out of 10,829) (see Table 3.4 and Table 3.5).

HISlink 2018

Similarly, in 2018, 5,692 households participated in the BHIS including in total 12,742 household members. Among them, 11,611 actually participated in the survey and 1,131 were not invited for participation. The NRNs of 12,731 individuals were found by Statbel, including 1 duplicate. Statbel sent the selected list of NRNs to IMA

via a TTP. At IMA level, 11,970 records were retrieved and 761 were not. Based on the MAF household composition in IMA DWH, 581 extra individuals were added leading to 12,551 records. In total, the HISlink 2018 contained 13,323 records (1 double) (see Figure 3.3). The overall linkage rate among individuals belonging to the participating households was 94.0% (11,970 out of 12,731). Among individuals who actually participated in the survey this percentage was 94.2% (10,933 out of 11,611) (see Table 3.6 and Table 3.7).

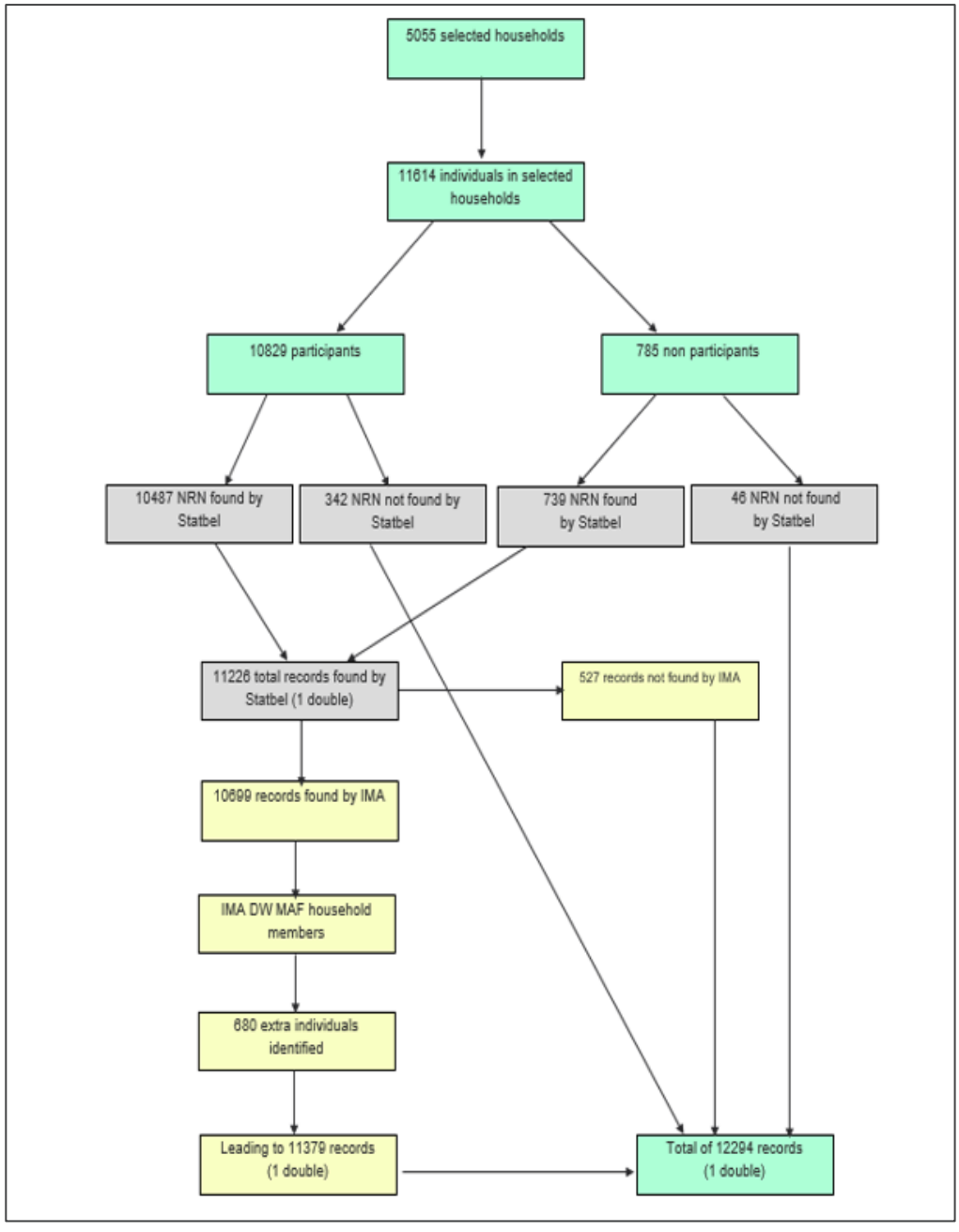


Figure 3.2: Data flow and overall result of the linkage, HISlink 2013, Belgium

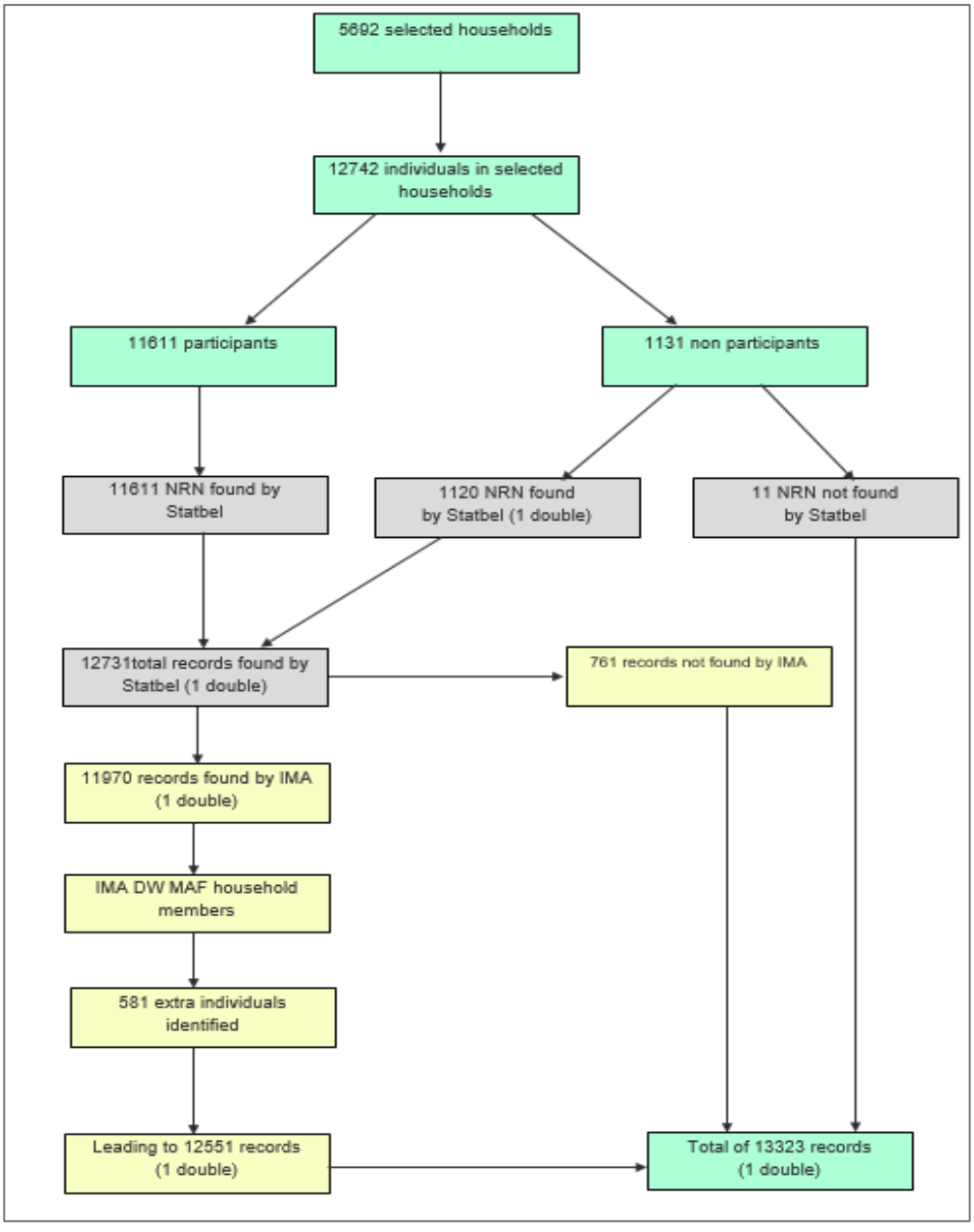


Figure 3.3: Data flow and overall result of the linkage, HISlink 2018, Belgium

Table 3.4: Distribution of individuals for whom data are available in the linked data file by type of data available - linkage status: individuals belonging to the participating households, HISlink 2013, Belgium

Linkage status of available data	**BHIS participants including people belonging to participating households but not selected to participate in the survey)**		**Final sample**	
	N	**%**	**N**	**%**
BHIS data linked with BCHI data	10699	92.1	10699	87.0
BHIS data not linked because Statbel could not find the NRN	388	3.3	388	3.2
BHIS data not linked because IMA could not find the NRN	527	4.6	527	4.3
Extra IMA data added because the head of MAF household/holder of MAF is part of BHIS sample	-	-	680	5.5
Total	**11614**	**100**	**12294**	**100**

Table 3.5: Distribution of individuals with data available in the linked data file by type of data available - linkage status: actual participants, Hislink 2013, Belgium

Linkage status of available data	**Actual participants to BHIS**		**People belonging to participating households but not selected to participate in the survey**	
	N	**%**	**N**	**%**
BHIS data linked with BCHI data	9998	92.3	701	89.3
BHIS data not linked because Statbel could not find the NRN	342	3.2	46	5.9
BHIS data not linked because IMA could not find the NRN	489	4.5	38	4.8
Total	**10829**	**100**	**785**	**100**

Table 3.6: Distribution of individuals for whom data are available in the linked data file by type of data available - linkage status: individuals belonging to the participating households, HISlink 2018, Belgium

Linkage status of available data	BHIS participants including people belonging to participating households but not selected to participate in the survey)		Final sample	
	N	%	N	%
BHIS data linked with BCHI data	11970	93.9	11970	89.8
BHIS data not linked because Statbel could not find the NRN	11	0.1	11	0.1
BHIS data not linked because IMA could not find the NRN	761	6.0	761	5.7
Extra IMA data added because the head of MAF household/holder of MAF is part of BHIS sample	-	-	581	4.4
Total	**12742**	**100**	**13323**	**100**

Table 3.7: Distribution of individuals with data available in the linked data file by type of data available - linkage status: actual participants Hislink 2018, Belgium

Linkage status of available data	Actual participants to BHIS		People belonging to participating households but not selected to participate in the survey	
	N	%	N	%
BHIS data linked with BCHI data	10933	94.2	1037	91.7
BHIS data not linked because Statbel could not find the NRN	-	0.0	11	1.0
BHIS data not linked because IMA could not find the NRN	678	5.8	83	7.3
Total	**11611**	**100**	**1131**	**100**

3.2.8. Quality evaluation and validation of the linked data

As described in chapter 2, several methods can be used to appraise the quality of linked data, including comparison of the characteristics between linked and unlinked data. Therefore, in order to identify particular subgroups of records that may be subject to linkage errors, the characteristics of linked and unlinked records are compared (16). Comparisons were performed using standardised differences, with a value greater than 0.1 indicating meaningful differences between groups (17–21). Such comparison helps to identify variables that may have been more affected by linkage error and are therefore potential sources of bias. The linkage rate was also calculated by subgroups.

Compared with characteristics of those whose data were linked, individuals whose data were not linked were more likely younger (0-14 years: 27.0% vs. 17.2%; standardised difference, -0.24), higher educated (55.7% vs 43.8%; standardised difference, -0.24), from not well-defined household composition, i.e. complex household or unknown category (17.8% vs. 8.6; standardised difference, -0.27) and foreigners (29.1% vs. 4.9%; standardised difference, -0.68 and 10.9% vs. 3.6%; standardised difference, -0.28, for Non-Belgian-EU and Non-Belgian-non EU, respectively). Records with linkage errors were also more likely to be from the Brussels Capital Region (29.6% vs. 10.2; standardised difference, -0.50) and Wallonia (38.4% vs. 31.9%; standardised difference, -0.14) but less likely to have a higher household income (13.4% vs. 18.7%; standardised difference, 0.14 and 12.8% vs. 20.5%; standardised difference, 0.21, for Quintile 3 and Quintile 4, respectively) for the HISlink 2013 (Table 3.8).

Similarly, in HISlink 2018, compared with linked data, unlinked data were more likely to be from male, 45-54 years, higher educated, Non-Belgian-EU, higher household income, the Brussels-Capital region and Wallonia (Table 3.9).

Several methods were used in the process of validating the linked data, including checking for consistency between the two data sources, identifying implausible values, comparing linkage rates, missed-match rates with previous linkages and comparing prevalence for a selection of indicators with previous reports.

Table 3.8: Comparison characteristics of the study population with linked and unlinked data, HISlink 2013, Belgium

Characteristics	Total N= 10829	Linked N=9998	Unlinked N=831	St. diff.	Linkage rate (%)
Gender, n (%)					
Male	5231 (48.7)	4819 (48.8)	412 (47.4)	0.03	92.1
Female	5598 (51.3)	5179 (51.2)	419 (52.6)	-0.03	92.5
Age, n (%)					
0-14	1716 (17.7)	1523 (17.2)	193 (27.0)	-0.24	88.7
15-24	1151 (11.6)	1051 (11.5)	100 (13.1)	-0.05	91.3
25-34	1406 (12.4)	1272 (12.3)	134 (15.3)	-0.08	90.5
35-44	1522 (13.8)	1378 (13.7)	144 (14.8)	-0.03	90.5
45-54	1558 (14.8)	1445 (14.9)	113 (13.3)	0.05	92.7
55-64	1450 (12.3)	1379 (12.5)	71 (7.5)	0.17	95.1
65-74	1032 (8.7)	998 (9.0)	34 (3.7)	0.21	96.7
75+	994 (8.7)	952 (8.9)	42 (5.3)	0.14	95.8
Education, n (%)					
Primary/No diploma	1133 (9.4)	1054 (9.4)	79 (9.1)	0.01	93.0
Lower secondary	1453 (12.2)	1389 (12.3)	64 (8.9)	0.11	95.6
Upper secondary	3395 (32.9)	3194 (33.3)	201 (24.6)	0.19	94.1
Higher education	4679 (44.3)	4211 (43.8)	468 (55.7)	-0.24	90.0
Missing	169 (1.2)	150 (1.2)	19 (1.7)	-0.05	88.7
Household composition, n (%)					
Single	1763 (14.9)	1685 (15.1)	78 (10.9)	0.12	95.6
One parent with child(ren)	1202 (9.0)	1115 (9.0)	87 (7.8)	0.04	92.8
Couple without child(ren)	2328 (21.9)	2203 (22.1)	125 (17.6)	0.11	94.6

Couple with child(ren)	4479 (45.2)	4105 (45.2)	374 (45.9)	-0.01	91.6
Other or unknown	1057 (9.0)	890 (8.6)	167 (17.8)	-0.27	84.2
Nationality, *n (%)*					
Belgian	9291 (89.9)	8834 (91.4)	457 (60.0)	0.79	95.1
Non-Belgian - EU	976 (6.1)	700 (4.9)	276 (29.1)	-0.68	71.7
Non-Belgian - non EU	555 (3.9)	457 (3.6)	98 (10.9)	-0.28	82.3
Missing	7 (0.1)	7 (0.1)	0 (-)	-	100
Household income, *n (%)*					
Quintile 1	2124 (17.1)	1983 (16.9)	141 (21.3)	-0.11	93.4
Quintile 2	1573 (14.7)	1516 (14.9)	57 (10.9)	0.12	96.4
Quintile 3	1841 (18.4)	1748 (18.7)	93 (13.4)	0.14	94.9
Quintile 4	1851 (20.2)	1768 (20.5)	83 (12.8)	0.21	95.5
Quintile 5	1974 (19.8)	1781 (19.7)	193 (22.1)	-0.06	90.2
Missing	1466 (9.7)	1202 (9.3)	264 (19.4)	-0.29	82.0
Region of residence, n (%)					
Flanders	3512 (56.7)	3425 (57.9)	87 (32.0)	0.54	97.5
Brussels	3103 (11.1)	2715 (10.2)	388 (29.6)	-0.50	87.5
Wallonia	4214 (32.2)	3858 (31.9)	356 (38.4)	-0.14	91.5

St. diff, standardised differences

Table 3.9: Comparison characteristics of the study population with linked and unlinked data, HISlink 2018, Belgium

Characteristics	Total N=11611	Linked N=10933	Unlinked N=678	St. diff.	Linkage rate (%)
Gender, n (%)					
Male	5588 (49.2)	5235 (49.0)	353 (56.7)	-0.16	93.7
Female	6023 (50.8)	5698 (51.0)	325 (43.3)	0.16	94.6
Age, n (%)					
0-14	1858 (17.6)	1766 (17.7)	92 (13.6)	0.11	95.0
15-24	1059 (11.3)	994 (11.3)	65 (11.1)	0.01	93.8
25-34	1338 (12.9)	1254 (12.8)	84 (15.7)	-0.08	93.7
35-44	1578 (12.7)	1461 (12.7)	117 (12.9)	-0.01	92.6
45-54	1725 (14.0)	1569 (13.8)	156 (21.2)	-0.19	90.9
55-64	1670 (13.1)	1584 (13.1)	86 (14.8)	-0.05	94.8
65-74	1289 (9.5)	1249 (9.7)	40 (5.4)	0.16	96.9
75+	1094 (8.8)	1056 (8.9)	38 (5.3)	0.14	96.5
Education, n (%)					
Primary/No diploma	811 (5.8)	779 (5.8)	32 (5.1)	0.03	96.0
Lower secondary	1434 (12.0)	1391 (12.2)	43 (6.9)	0.18	97.0
Upper secondary	3402 (31.6)	3279 (32.0)	123 (18.9)	0.30	96.4
Higher education	5755 (49.3)	5309 (48.8)	446 (67.3)	-0.38	92.2
Missing	209 (1.3)	175 (1.2)	34 (1.8)	-0.04	83.7
Household composition, n (%)					
Single	2151 (15.4)	2047 (15.5)	104 (14.1)	0.04	95.2
One parent with child(ren)	1276 (10.8)	1228 (10.9)	48 (6.6)	0.15	96.2
Couple without child(ren)	2598 (22.5)	2469 (22.4)	129 (24.2)	-0.04	95.0

Couple with child(ren)	5017 (46.4)	4656 (46.3)	361 (51.5)	-0.11	92.8
Other or unknown	569 (4.9)	533 (4.9)	36 (3.6)	0.07	93.7
Nationality, *n (%)*					
Belgian	9761 (88.9)	9461 (90.1)	300 (50.1)	0.97	96.9
Non-Belgian - EU	1184 (6.3)	846 (5.2)	338 (43.0)	-0.98	71.4
Non-Belgian – non-EU	661 (4.7)	621 (4.7)	40 (6.9)	-0.09	93.9
Missing	5 (0.1)	5 (0.1)	0 (-)	-	100
Household income, *n (%)*					
Quintile 1	1221 (8.8)	1192 (8.9)	29 (6.0)	0.11	97.6
Quintile 2	1476 (11.7)	1450 (11.9)	26 (3.2)	0.33	98.2
Quintile 3	1861 (16.2)	1820 (16.5)	41 (8.3)	0.25	97.8
Quintile 4	2406 (22.1)	2322 (22.4)	84 (11.8)	0.28	96.5
Quintile 5	2804 (27.0)	2487 (26.4)	317 (47.5)	-0.45	88.7
Missing	1843 (14.2)	1662 (13.9)	181 (23.2)	-0.24	90.2
Region of residence, n (%)					
Flanders	4296 (55.8)	4230 (56.5)	66 (31.4)	0.52	98.5
Brussels	3099 (10.6)	2873 (10.2)	226 (26.2)	-0.42	92.7
Wallonia	4216 (33.6)	3830 (33.3)	386 (42.4)	-0.19	90.8

St. diff, standardised differences

3.2.9. Timing of the linkage procedure

Tables 3.10 and 3.11 show the main steps in the linkage procedure and the time needed for each step for HISlink 2013 and HISlink 2018 respectively. In addition to the preparatory tasks (e.g. meetings, development of the linkage scheme, drafting of the authorisation request, data preparation, etc.) which can take several months, up to 17 and 18 months elapsed between the submission of the authorisation request and the publication of the linked data for HISlink 2013 and HISlink 2018, respectively. As shown in Table 3.11, the IRB procedure took longer for HISlink 2018 than for HISlink 2013. This could be explained by the fact that, although the GDPR had not yet been fully implemented at the time of the application, the IRB had become more stringent than it was in 2017. In addition, the actual linkage took longer for HISlink 2018 due to Covid-19 and the mobilisation of resources for the related projects. Finally, there were two data deliveries for HISlink 2018. The first in September 2021, when an error was found after the first explorations, due to the use of the wrong database in the linkage process. The correctly linked databases were made available in December 2021.

In view of the time required for the entire linkage process, from the preparatory stages to the delivery of the linked data, it is important to take the lead and begin the preparation, including the administrative stage (submission of the request to ISC) in parallel with the BHIS fieldwork, which takes around 1 year. In this way, the actual linkage could take place as soon as the BHIS data are ready. This proactive strategy, which should enable the linked data to be made available more quickly after the survey, is currently being tested for HISlink 2023.

Table 3.10: Timeline of the main steps of linkage procedure, HISlink 2013

Activities	2017			2018													2019		
	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M
Submission of authorisation request																		
Publication of deliberation authorisation request																		
TTP eHealth Global agreement doc																		
Small Cell Risk Analysis by IMA																		
Preparation of BHIS data to be sent to Statbel																		
Transferring BHIS data to Statbel																		
Linkage procedure at the level of health insurance funds, IMA and TTP																		
Linked data made available to Sciensano researchers on IMA server																		

Table 3.11: Timeline of the main steps of linkage procedure, HISlink 2018

Activities	2020							2021										
	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D
Submission of authorisation request																			
Publication of deliberation authorisation request																			
Small Cell Risk Analysis by IMA																			
TTP eHealth Global agreement doc																			
Preparation of BHIS data to be sent to Statbel																			
Transferring BHIS data to Statbel																			
Linkage procedure at the level of health insurance funds, IMA and TTP																			
Linked data made available to Sciensano researchers on IMA server (first release)																			
Linked data made available to Sciensano researchers on IMA server (second release after correction)																			

3.3. BIBLIOGRAPHY

1. Sciensano. HIS - Health Interview Survey [Internet]. [cited 2023 May 25]. Available from: https://www.sciensano.be/en/projects/health-interview-survey
2. Demarest S, Van der Heyden J, Charafeddine R, Drieskens S, Gisle L, Tafforeau J. Methodological basics and evolution of the Belgian health interview survey 1997–2008. Arch Public Health. 2013 Dec;71(1):24.
3. de Leeuw ED. Internet Surveys as Part of a Mixed-Mode Design. In: Das M, Ester P, Kaczmirek L, editors. Social and Behavioral Research and the Internet [Internet]. 1st ed. Routledge; 2018 [cited 2023 May 29]. p. 45–76. Available from: https://www.taylorfrancis.com/books/9781136923586/chapters/10.4324/9780203844922-3
4. De Leeuw ED. To mix or not to mix data collection modes in surveys. Journal of official statistics. 2005;21(5):233–55.
5. Ombelet F, Goossens E, Willems R, Annemans L, Budts W, De Backer J, et al. Creating the BELgian COngenital heart disease database combining administrative and clinical data (BELCODAC): Rationale, design and methodology. International Journal of Cardiology. 2020 Oct;316:72–8.
6. Bouckaert N, Maertens de Noordhout C, Van de Voorde C. Health System Performance Assessment: how equitable is the Belgian health system? [Internet]. Brussels: Belgian: Health Services Research (HSR). Health Care Knowledge Centre (KCE); 2020 [cited 2022 Jun 27] p. 105. Report No.: KCE Reports 334. D/2020/10.273/30. Available from: https://kce.fgov.be/sites/default/files/2021-11/KCE_334_Equity_Belgian_health_system_Report.pdf
7. Agence InterMutualiste -InterMutualistisch Agentschap (AIM-IMA). Agence InterMutualiste -InterMutualistisch Agentschap [Internet]. [cited 2021 Jul 26]. Available from: https://www.ima-aim.be/-Donnees-de-sante-
8. Vaes B, Ruelens C, Saikali S, Smets A, Henrard S, Renard F, et al. Estimating the prevalence of diabetes mellitus and thyroid disorders using medication data in Flanders, Belgium. European Journal of Public Health. 2018 Feb 1;28(1):193–8.
9. Agence InterMutualiste -InterMutualistisch Agentschap. Metadata [Internet]. [cited 2023 Jun 29]. Available from: https://metadata.ima-aim.be/fr/app/vars/FKG_XXX_Pa
10. Maetens A, De Schreye R, Faes K, Houttekier D, Deliens L, Gielen B, et al. Using linked administrative and disease-specific databases to study end-of-life care at population level. BMC Palliat Care. 2016 Dec;15(1):86.
11. Larcin L, Lona M, Karakaya G, Van Espen A, Damase-Michel C, Kirakoya-Samadoulougou F. Using administrative healthcare database records to study trends in prescribed medication dispensed during pregnancy in Belgium from 2003 to 2017. Pharmacoepidemiol Drug Saf. 2021 Sep;30(9):1224–32.

12. Charafeddine R, Berger N, Demarest S, Van Oyen H. Using mortality follow-up of surveys to estimate social inequalities in healthy life years. Popul Health Metrics. 2014 Dec;12(1):13.
13. Yokota RTC, Nusselder WJ, Robine JM, Tafforeau J, Charafeddine R, Gisle L, et al. Contribution of chronic conditions to smoking differences in life expectancy with and without disability in Belgium. European Journal of Public Health. 2018 Oct 1;28(5):859–63.
14. Van der Heyden J, De Bacquer D, Tafforeau J, Van Herck K. Reliability and validity of a global question on self-reported chronic morbidity. J Public Health. 2014 Aug;22(4):371–80.
15. Mimilidis Hélène, Demarest Stefaan, Tafforeau Jean, Van der Heyden Johan. Projet de couplage de données issues de l'Enquête de Santé 2008 et des Organismes Assureurs. Bruxelles, Belgique; 2014 Mai. Report No.: D/2014/2505/32.
16. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. International Journal of Epidemiology. 2017 Oct 1;46(5):1699–710.
17. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statist Med. 2009 Nov 10;28(25):3083–107.
18. Austin PC. Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research. Communications in Statistics - Simulation and Computation. 2009 May 14;38(6):1228–34.
19. Son M, Gallagher K, Lo JY, Lindgren E, Burris HH, Dysart K, et al. Coronavirus Disease 2019 (COVID-19) Pandemic and Pregnancy Outcomes in a U.S. Population. Obstetrics & Gynecology. 2021 Oct;138(4):542–51.
20. Pincus D, Jenkinson R, Paterson M, Leroux T, Ravi B. Association Between Surgical Approach and Major Surgical Complications in Patients Undergoing Total Hip Arthroplasty. JAMA. 2020 Mar 17;323(11):1070.
21. Bayoumi AM. STDDIFF: Stata module to compute standardized differences for continuous and categorical variables. 2021.

3.4. ANNEX

Table A1: Overview of the contents of the HISlink 2013 and 2018 databases, Belgium

Modules	**From BHIS source**	**From BCHI source**
	Description / Operationalisation	**Description / Operationalisation**
Information related to the survey		
ID	Identification number of respondent	
Participated in survey	Status of actual participation in the survey (yes/no)	
Date of interview	Date of interview (DD/MM/YYYY)	
Year of the survey	Year of the survey (YYYY)	
Weight of individual within the sample	Individual post stratification weights	
Availability of self-completed questionnaire	Status of self- competed questionnaire: - Self-competed questionnaire not required and not available - Self-competed questionnaire not required, but available -Self-completed questionnaire required, but not available -Self-completed questionnaire required and available	
Household cluster	Identification of the household	
Socio-demographic characteristics		
Age	Age (in years)	Year of birth
Sex	Gender (Male / Female)	Gender (Male / Female)
Education	Educational attainment based on the highest level of education achieved according to the ISCED 1997 in four categories: No diploma or primary education / Lower secondary / Higher secondary / Higher.	
Place of residence	Province (11 categories) / Region (3 categories) of residence at the time of the survey	Province (11 categories) of residence at the time of the survey
Household composition / number of persons in the household	Household composition based on reference person in national register.	Health insurance household based on MAF head of household.
Nationality/country of birth	Nationality / Country of birth (3 categories): Belgian / Non-Belgian – EU / Non-Belgian – non-EU.	
Housing	Housing tenure (Owner, co-owner or usufructuary, Renter from an individual private landlord or society, Renter from a social housing association or another public association, Living rent-free).	
Income	The equivalent household income (quintiles based on Belgian population)	

Employment	Current (last) employment / non-employment status	Unemployment status during the last trimester preceding the reference year.
Entitlement to increased reimbursement		Receipt of a disability or invalidity allowance, take-up and use of increased reimbursement status, maximum billing system, Lump sum for the chronically ill.
Insurance status		Insurance status of the individual: Undefined situations or no entitlement / Employee (under the general scheme) entitled to large risks / Self-employed person entitled to comprehensive cover".
Health status		
Perceived health	Self-reported indicator based on the question: "How is your health in general?". Five response categories are possible: Very good / Good / Fair / Poor / Very poor.	
Chronic conditions	Self-reported chronic conditions based on the question: "Have you suffered during the last 12 months from the following disease?" followed by a list of 35 chronic conditions: asthma, chronic bronchitis, chronic obstructive pulmonary disease or emphysema, Parkinson's disease, high blood pressure, epilepsy, myocardial infarction, coronary heart disease, serious heart disease (except myocardial infarction of coronary heart disease), stroke (or consequences), chronic fatigue for a period of at least 3 months, rheumatoid arthritis, osteoarthritis, osteoporosis, diabetes, disorder of the larger or the small bowel for at least 3 months, allergy, serious disease of the kidney other than stones in the kidney, stones in the kidney, stomach ulcer, chronic cystitis, cirrhosis of the liver, liver dysfunction, serious or chronic skin disease, cancer, gallstones or inflammation of the gallbladder, severe headache such as migraine, serious dejection or depression, thyroid problems, high cholesterol level in blood, narrowing of blood vessels, lower back disorder, neck disorder, urinary incontinence, broken hip, prostate problems, eye diseases (diabetic retinopathy, macular degeneration, cataract, glaucoma, other eye disease).	Proxy for diagnostic information (pseudo pathologies) based on the ATC-codes of dispensed medication in public pharmacies, including: cardiovascular disorders, diabetes, asthma, epilepsy, chronic obstructive pulmonary disease, thyroid disorders, cancers, Parkinson's disease, HIV, cystic fibrosis, exocrine pancreatic diseases, psoriasis, rheumatoid arthritis, psychosis, chronic hepatitis B and C, multiple sclerosis, organ transplantation, Alzheimer's disease, renal failure, haemophilia.
Functional limitations	Self-reported functional limitations / restrictions in daily activities due to health problems	
Mental health	Self-reported information on different dimensions of mental health: well-being/distress, disorders/symptoms, eating behaviours, suicidal behaviours, positive mental health/vitality, use of psychotropic medicine and self-perceived depression.	
Physical pain	Self-reported bodily pain during the past four weeks.	
Quality of life	Self-reported information on the impact of health status on quality of life, assessed along five dimensions: mobility, personal autonomy, daily activities, pain/discomfort and anxiety/depression.	
Absence from work because of health problems	Self-reported information on absenteeism and number of days absent from work due to health problems.	

Frailty*	Self-reported information on the vulnerability or fragility of the elderly population.	
Children's strengths and difficulties	Self-reported information on emotional, behavioural and attentional disorders in children and adolescents, and their possible management.	
Lifestyle and health behaviour		
Smoking	Self-reported information on smoking (current smokers, former smokers and non-smokers).	
Use of electronic cigarettes*	Self-reported use of e-cigarettes or similar devices such as electronic chicha, pipes or cigars	
Use of alcohol	Self-reported alcohol consumption.	
Use of illicit drugs	Self-reported use of illicit drugs.	
Physical activity	Self-reported physical activity.	
Nutritional status	Self-reported nutritional status (weight, height) and the resulting body mass index – BMI.	
Nutritional habits	Self-reported eating habits of the population: consumption of fruit, vegetables or salads, 100% pure juices, sweetened drinks, sweet or salty snacks, calcium-enriched dairy products or vegetable products, the amount of water drunk daily, etc. or vegetable products enriched with calcium, the amount of water drunk daily and breakfast (frequency); food allergies or intolerances doctor).	
Dental health	Self-reported information on oral health in the population: use of dental prostheses, frequency of brushing, frequency of tooth brushing, limitations caused by oral problems.	
Sexual health	Self-reported information on practices and use of different methods of contraception within the population.	
Gambling*	Self-reported information on gambling addiction problems (casino games, slots, bingo, scratch cards, sports betting, etc.).	
Health Prevention and attitudes		
Cancer screening	Self-reported information on colorectal cancer, breast cancer and cervical cancer screening, based on the question: "Have you ever had a faecal occult blood test / colonoscopy /mammography / cervical smear test?" and "When was the last time you had a faecal occult blood test / colonoscopy /mammography / cervical smear test?".	Specific nomenclature codes for reimbursement of mammograms performed as part of screening programmes or outside screening programmes.
Vaccination	Self-reported vaccination against influenza, pneumococcus and human papillomavirus.	ATC codes of supplied vaccines
Screening for cardiovascular risk factors and diabetes	Self-reported information on methods of preventing cardiovascular disease and diabetes, including control of blood pressure, blood sugar and cholesterol levels.	
Knowledge and attitudes towards HIV	Self-reported information on the knowledge and beliefs of the population on the transmission of the AIDS virus and effective methods of protection against transmission.	

Health literacy*	Self-reported information on the level of health literacy in the population (motivation and skills of individuals to access, understand, evaluate and use information to make decisions about their health).	
Use of health care and other services		
Contacts with GP	Self-reported information on contacts with GP in the last 12 months.	Specific nomenclature and competence codes for contacts with GPs in outpatient and inpatient settings, date and type of services.
Contacts with specialist	Self-reported information on contacts with specialist in the last 12 months.	Specific nomenclature and competence codes for contacts with specialists in outpatient and inpatient settings, date and type of services.
Contacts with emergency department of hospital	Self-reported information on contacts with emergency department of hospital in the last 12 months.	Specific nomenclature codes for contacts with emergency department of hospital, date and type of services
Contacts with dentist	Self-reported information on contacts with dentist in the last 12 months.	Specific nomenclature and competence codes for contacts with dentists, date and type of services.
Contact with paramedical professionals	Self-reported information on contacts with paramedical professionals in the last 12 months.	Specific nomenclature and competence codes for contacts with paramedical professionals in outpatient and inpatient settings, date and type of services
Contact with practitioners of non-conventional medicine	Self-reported information on contacts with practitioners of non-conventional medicine in the last 12 months.	
Contacts with home-care services	Self-reported information on the use of home-care services in the last 12 months in the event of health problems. These services comprise for example home-care services provided by a nurse or midwife, home help for housework or for older people, "meals on wheels" or transport service.	
Admission to hospital	Self-reported information on hospital admission in the last 12 months.	Specific codes for outpatient and inpatient admission, admission date, discharge date, types of services /procedures.
Admission to rest home or nursing home		Nomenclature number, service start date, amount (co-payment + reimbursed).
Use of medicines	Information on the actual use of medicines, including information on the medicines prescribed or purchased. Thus, both non-prescription and prescription medicines are also taken into account. The definition of "medicine" is broader and includes food supplements, medicinal plants, homeopathic products, contraceptive pills, etc.	Prescriptions for reimbursable medicines: CNK code, date of supply, amount (co-payment + reimbursed)
Accessibility of health care	Self-reported relative burden of healthcare expenditure on household budgets.	
Patient experiences	Self-reported information on patients' experiences from the moment they make an appointment to the consultation and prescription of treatment, in order to obtain an overall assessment of the quality of care services.	

Physical and social health environment		
Passive smoking	Self-reported passive smoking.	
Other environmental factors affecting health	Self-reported information on the nuisance in the neighbourhood or district and the nuisance felt at home (inside the house) and coming from the immediate environment. home (inside the house) and from the immediate environment.	
Accidents	Self-reported information on domestic, road or leisure accidents during leisure time, resulting in injury as well as falls among older population.	
Violence	Self-reported information on interpersonal violence.	
Social health	Self-reported information on integration into a social network and the support that the person can have in the event of a problem. This information involves identifying groups of people who are isolated or socially deprived and examining the link with physical and mental health.	
Informal care	Self-reported information on informal (non-professional) help given to people with age-related problems or long-term illnesses. The perspective is that of providing help, not receiving it.	
Mortality data		
Death status		Status of death (Yes /No)
Date of death		Date of death (dd/mm/yyyy)
Healthcare expenditures		
Amount reimbursed for healthcare use		Amount refunded by health insurance.
Out-of-pocket		Personal intervention.
Supplements		Additional amount or amount for non-refundable products, services or supplies.

**Available in HISlink 2018 only.*

CHAPTER 4. USE OF LINKED DATA AS VALIDATION TOOL

4.1. VALIDITY OF SELF-REPORTED MAMMOGRAPHY UPTAKE IN THE BELGIAN HEALTH INTERVIEW SURVEY: SELECTION AND REPORTING BIAS

The findings of this paper were published as:

Berete F, Van der Heyden J, Demarest S, Charafeddine R, Tafforeau J, Van Oyen H, Bruyère O and Renard F. Validity of self-reported mammography uptake in the Belgian health interview survey: selection and reporting bias. European Journal of Public Health 31.1 (2021): 214-220.

European Journal of Public Health, 1–7

doi:10.1093/eurpub/ckaa217

Validity of self-reported mammography uptake in the Belgian health interview survey: selection and reporting bias

Finaba Berete[1,2]*, Johan Van der Heyden[1], Stefaan Demarest[1], Rana Charafeddine[1], Jean Tafforeau[1], Herman Van Oyen[1,3], Olivier Bruyère[4], Françoise Renard[1]

1 Department Epidemiology and Public Health, Sciensano, Brussels, Belgium
2 Department of Public Health, Epidemiology and Health Economics, University of Liège, Liège, Belgium
3 Department of Public Health and Primary Care, Ghent University, Ghent, Belgium
4 Department of Public Health, Epidemiology and Health Economics, WHO Collaborating Centre for Public Health Aspects of Musculoskeletal Health and Ageing, University of Liege, Liège, Belgium

Correspondence: F. Berete, SD Epidemiology and Public Health, Sciensano, Juliette Wytsmanstraat 14, 1050 Brussels, Belgium, Tel: +32 2 642 54 76, Fax: + 32 2 642 50 01, e-mail: Finaba.berete@sciensano.be

Background: The validity of self-reported mammography uptake is often questioned. We assessed the related selection and reporting biases among women aged 50–69 years in the Belgian Health Interview Survey (BHIS) using reimbursement data for mammography stemming from the Belgian Compulsory Health Insurance organizations (BCHI). **Methods:** Individual BHIS 2013 data ($n = 1040$) were linked to BCHI data 2010–13 (BHIS–BCHI sample). Being reimbursed for mammography within the last 2-years was used as the gold standard. Selection bias was assessed by comparing BHIS estimates reimbursement rates in BHIS–BCHI with similar estimates from the Echantillon Permanent/Permanente Steekproef (EPS), a random sample of BCHI data, while reporting bias was investigated by comparing self-reported versus reimbursement information in the BHIS–BCHI. Reporting bias was further explored through measures of agreement and logistic regression. **Results:** Mammography uptake rates based on self-reported information and reimbursement from the BHIS–BCHI were 75.5% and 69.8%, respectively. In the EPS, it was 64.1%. The validity is significantly affected by both selection bias {relative size = 8.93% [95% confidence interval (CI): 3.21–14.64]} and reporting bias [relative size = 8.22% (95% CI: 0.76–15.68)]. Sensitivity was excellent (93.7%), while the specificity was fair (66.4%). The agreement was moderate (kappa = 0.63). Women born in non-EU countries (OR = 2.81, 95% CI: 1.54–5.13), with high household income (OR = 1.27, 95% CI: 1.02–1.60) and those reporting poor perceived health (OR = 1.41, 95% CI: 1.14–1.73) were more likely to inaccurately report their mammography uptake. **Conclusions:** The validity of self-reported mammography uptake in women aged 50–69 years is affected by both selection and reporting bias. Both administrative and survey data are complementary when assessing mammography uptake.

4.1.1. Abstract

Background

The validity of self-reported mammography uptake is often questioned. We assessed the related selection and reporting biases among women aged 50-69 years in the Belgian Health Interview Survey (BHIS) using reimbursement data for mammography stemming from the Belgian Compulsory Health Insurance organizations (BCHI).

Methods

Individual BHIS 2013 data (n=1,040) were linked to BCHI data 2010-2013 (BHIS-BCHI sample). Being reimbursed for mammography within the last 2-years was used as the gold standard. Selection bias was assessed by comparing BHIS estimates reimbursement rates in BHIS-BCHI with similar estimates from the Echantillon Permanent/Permanente Steekproef (EPS), a random sample of BCHI data, whilst reporting bias was investigated by comparing self-reported versus reimbursement information in the BHIS-BCHI. Reporting bias was further explored through measures of agreement and logistic regression.

Results

Mammography uptake rates based on self-reported information and reimbursement from the BHIS-BCHI were 75.5% and 69.8%, respectively. In the EPS it was 64.1%. The validity is significantly affected by both selection bias (relative size=8.93% (95% CI:3.21-14.64)) and reporting bias (relative size=8.22% (95% CI: 0.76-15.68)). Sensitivity was excellent (93.7%) while the specificity was fair (66.4%). The agreement was moderate (kappa=0.63). Women born in non-EU countries (OR=2.81, 95% CI:1.54-5.13), with high household income (OR=1.27, 95% CI:1.02-1.60) and those reporting poor perceived health (OR=1.41, 95% CI:1.14-1.73) were more likely to inaccurately report their mammography uptake.

Conclusions

The validity of self-reported mammography uptake in women aged 50-69 years is affected by both selection and reporting bias. Both administrative and survey data are complementary when assessing mammography uptake.

Keywords: Validity, mammography uptake, selection bias, reporting bias, breast cancer screening, data linkage.

4.1.2. Introduction

Breast cancer is the most common cancer in terms of incidence among women both in developed and developing countries [1-3] and the second cause of cancer death among women after lung cancer in most developed countries [4].

Early detection of breast cancer through mammography screening is recognized as being effective in reducing mortality [5;6] in women aged 50 to 69 years [7-9]. Literature suggests that with a screening attendance reaching 70%, a reduction in breast cancer mortality by about 25% might be expected [8;10]. European guidelines recommend biennial mammography screening for women aged 50-69 years [11].

Valid methods of determining and monitoring breast cancer screening (screening) uptake are important to evaluate screening programs [6;12;13]. Underestimating screening prevalence could lead to waste of resources, while overestimation could lead to missed opportunities for improving screening [6].

Currently, information on screening is often based on self-reports in population-based surveys [5;6;12;13]. Such information is used to monitor screening rates over time and to target interventions. However, the validity of self-reported information through surveys is a concern due to a potential selection (because of non-coverage or non-response error) and reporting bias associated with differential survey participation. Survey participants may systematically differ from the general population and reporting may be inaccurate due to memory and social desirability effects. The validity can also be different for different subpopulations. E.g. it has been shown that members of ethnic minority groups and people with a lower socioeconomic status are more likely to inaccurately report cancer screening than their counterparts [5;13;14]

According to the European screening quality assurance guidelines, the acceptable and desirable participation rates of screening are 70% and 75% respectively [9]. Furthermore, the European Partnership for Action Against Cancer called for reducing the burden of cancer by achieving 100% population coverage of screening for breast, cervical and colorectal cancer in 2013 [15;16]. In the US, the Healthy People 2020 goals calls for a rate of adherence to national cancer-screening guidelines of 81% biannual mammography among women aged 50-74 years [17;18].

To verify whether these goals are met, it is necessary to ensure that the data for estimating the national mammography uptake rate are valid.

The validity of self-reported mammography uptake can be verified by comparing this information with a trusted measure (gold standard). Numerous validation studies and meta-analyses have documented the level of agreement/disagreement between self-reported cancer screening and cancer registers, claims databases, electronic medical records and administrative data [5;6;12;13]. They have reported a sensitivity between 95% to 97%, a specificity between 61% to 64% and have concluded that the estimates based on self-report are usually over-estimated [5;12]. However, most of these validation studies are either limited to a specific geographical region [6;19-21] or a specific subgroup [13;14], leading to a problem of generalizability.

In Belgium, breast cancer is the first female cancer in terms of incidence (more than a third of cancers) [22] and the leading cause of premature death among women [23].

A national mammography screening program exists in Belgium since 2001-2002. Mammograms realized within this organized screening program are called “mammotests”. Such mammograms are entirely reimbursed by the National Institute of Health and Disability Insurance (NIHDI) as well as diagnostic mammograms (i.e. among symptomatic women or those at high-risk). The mammotests and diagnostic mammograms are coded differently in the BCHI database. Besides the mammotests, a number of screening was often realized by women outside of the official screening program (by their own initiative). These later are called “opportunistic screening” and are not reimbursed by the NIHDI. More often, for the reimbursement purposes, the opportunistic screening are miscoded as diagnostic mammograms. The proportion of mammograms realized outside of the screening program is important. Thus, information on mammography uptake gathered through the BCHI data allow to capture the total coverage of the screening than those through the official screening program. Each woman aged 50-69 years receives every 2 years an invitation to participate in the screening program. The mammograms realized within the program follow a specific procedure. The examination is free of charge [24]. Exhaustive information on the mammography uptake is available through the Belgian Compulsory Health Insurance (BCHI) including both mammograms realized within and outside the organized screening program [25]. However, the BCHI database is limited in terms of sociodemographic information.

Information on mammography uptake ("having had mammograms"), based on self-reports is available in the Belgian Health Interview Survey (BHIS) [26]. The added value of the BHIS data is that it provides a comprehensive information on socioeconomic status (SES) and many other health related topics useful for subgroup analyzes. Nevertheless, as in other population surveys, selection and reporting bias are also a concern in the BHIS [27]. In this study, we investigate the validity of self-reported mammography attendance in the BHIS, as a proxy of screening uptake by assessing the associated selection and reporting biases.

4.1.3. Methods

Data sources

BHIS 2013 data were linked to BCHI 2010-2013 data (BHIS-BCHI) by means of a unique identifier (the national register number). The BHIS is a national, cross-sectional household survey conducted every 5 years since 1997 by Sciensano among a representative sample of Belgian residents. Participants are selected from the national population register through a multistage stratified sampling procedure. The detailed methodology of the survey is described elsewhere [28].

The BHIS collects information on mammography uptake by means of a self-administered questionnaire in women aged 15 years and older (the reference population of Eurostat, although the main indicator refers to women aged 50-69 years): Have you ever had mammograms? "Yes/No" and for those who respond "Yes", the time lapse since her last mammograms: "When was the last time you had mammograms? Furthermore, the BHIS also collects data on a wide range of other health and health related topics such as demographic information, SES and self-reported health status, life style and health services use. The BHIS has been approved by the ethics committee of the University hospital of Ghent on October, 1st 2012 (advice EC UZG 2012/658). For the linkage, an authorization was obtained from the Belgian Privacy Commission.

BCHI data contain exhaustive and detailed information on the reimbursed health expenses of over 99% of the total population. The database also includes a limited amount of socio-demographic information [29].

The Echantillon Permanent/Permanente Steekproef (EPS) data, representing 1/40 of the Belgian population, is an unbiased random sample of the BCHI and contains the same information as the BCHI (reimbursed medical acts, hospitalization, and medicines) which is also followed over time. The use the data of this population cohort in an anonymized way for policy and research purposes is regulated by a specific legal framework [27]. All women aged 50-69 years within the EPS are included in this study. The analysis of the EPS data does not require any design settings.

Inclusion criteria

This study included women aged 50-69 years who responded to the questions related to the mammography uptake "having had mammograms" of BHIS (n=1,081). Linkage with BCHI was possible for 1,040 women (96%). To assess the validity of self-reported mammography uptake, reimbursement for a mammography within the last 2 years preceding the BHIS was used as the gold standard. As nor in BHIS nor in BCHI it is not possible to disentangle mammograms realized within the screening program from opportunistic screening, both types are included in this study. We assume that in both sources, the mammography uptake in this age group is a good proxy for the screening uptake.

Analyses

Mammography uptake rates by data source were calculated.

Selection bias

The potential selection bias was computed as the difference between the prevalence of register within BCHI based mammography uptake from the BHIS-BCHI and similar estimates from the EPS data (absolute bias), and dividing that difference by the prevalence from the EPS data and multiplied by 100 (relative bias) [30]. The 95% CI of the estimated bias were computed using the Delta method [29]. Analyses were done overall and by age-group and region of residence.

Reporting bias

The reporting bias was assessed as the difference in the prevalence of mammography uptake between BHIS and BCHI estimates from the BHIS-BCHI linked data. As for the selection bias both absolute and relative percentages were calculated. Next, the report-to-record ratio (RRR) was computed . The RRR is the ratio of the percentage of women reporting having had mammograms to the percentage of women

reimbursed for mammograms during the relevant time period, and its confidence intervals. The RRR is frequently used as a measure of net bias of self-report, with values greater than one indicating over-reporting and values less than one indicating under-reporting [6;13;31;32]. Furthermore the sensitivity (i.e., the percentage of women classified as screened in the BHIS among those who were reimbursed for mammograms in the BCHI, the specificity (i.e. the percentage of women classified as not screened in the BHIS, among those who were not reimbursed for mammograms in the BCHI), the positive predictive value (PPV, i.e. the percentage of women reimbursed for mammograms in the BCHI, among those classified as screened in the BHIS, the negative predictive value (NPV, i.e. the percentage of women who were not reimbursed for mammograms in the BCHI, among those who were classified as not screened in the BHIS were calculated. These estimates were classified as excellent (>0.90), good (>0.80), fair (>0.70), or poor (<0.70) [33]. Sensitivity analysis was performed by moving the time frame for screening from 2 to 3 years. The total agreement as well as the Cohen's kappa statistic were also calculated to provide a measure of agreement beyond chance [34]. Cutoffs used to classify kappa are based on McHugh et al. : 0-0.20 = none agreement; 0.21-0.39 = minimal agreement; 0.40- 0.59 = weak agreement; 0.60-0.79 = moderate agreement; 0.80-0.90 = strong agreement; above 0.90 = almost perfect agreement [35].

The calculations were done for the whole population and by sociodemographic subgroups; by age-group (50-59 years, 60-69 years), educational level, country of birth (Belgium, other EU country, non-EU country), region of residence (Flanders, Brussels and Wallonia), income category (low, high) and self-perceived health (good to very good, very bad to fair). Educational level was based on the highest level of education achieved in the household according to the ISCED 1997 [36] and recoded unto three categories: low (lower secondary education or less), intermediate (higher secondary education), and high (higher education). For income level, the quintiles of the equivalent household income were recorded in low (quintile 1 to 3) and high (quintile 4 and 5). As this is an exploratory and post-hoc analysis of existing data, a strict adjustment for multiple comparison is less critical [37]. Therefore, we declined to adjust for multiple comparisons.

Finally, multivariable logistic regression was used to identify covariates associated with inaccurate self-reported mammography uptake (over- or underreporting). All

variables cited above were included as independent variables. In order to maximize the information available in the analyses and to prevent potential bias caused by selective drop out, item nonresponse for education, household income, perceived health and place of birth (item missingness between 1% to 10%) was addressed by multiple imputation. Age, region of residence, as well as the dependent variable were used in the imputation model. The dependent variable (3.5% of missingness) was included in the imputation model in order to enhance it and was reliably imputed. However its imputed values were not used in the analysis model. Multivariate normal regression was used as the imputation method to estimate missing values [38]. Survey data were analyzed taking into account the multistage stratified clustered sampling design of the BHIS: use of post stratification weights, geographical stratification at the level of the province and clustering at household level. Statistical significance was defined as $P < 0.05$. Potential selection and reporting bias were estimated using Stata 15.1©. All the remaining analyses were performed using SAS 9.4©.

4.1.4. Results

Table 4.1.1 presents the prevalence of mammography uptake by data source and subgroups. Based on the BHIS-BCHI, the mammography uptake in the BHIS 2013 sample was estimated to be 75.5% using BHIS information and 69.8% using BCHI information. Within the EPS, the percentage was 64.1%. The percentage also varies significantly across subgroups in both data source.

Table 4.1 1: Prevalence of mammography uptake in the last 2 years, by source and subgroups, HISlink 2013, Belgium

	BHIS-BCHI linked (n = 1,040)				EPS (n =36,700)	
	BHIS		BCHI			
	% uptake	95% CI	% uptake	95% CI	% uptake	95% CI
Overall	75.5	(72.1-78.9)	69.8	(66.2-73.4)	64.1	(63.6-64.6)
Age (years)						
50-59	78.0	(73.6-82.4)	68.6	(63.6-73.6)	67.1	(66.4-67.7)
60-69	72.8	(67.6-77.9)	71.1	(65.8-76.3)	67.0	(66.3-67.8)
Educational level					*N.A*	
Low	66.2	(59.4-73.1)	61.0	(53.5-68.4)		
Middle	76.4	(70.3-82.4)	73.4	(67.1-79.6)		
High	81.7	(76.9-86.4)	73.7	(68.2-79.2)		
Place of birth					*N.A*	
Belgium	76.0	(72.4-79.5)	70.1	(66.3-73.9)		
EU country	63.2	(50.1-76.3)	57.5	(44.0-71.1)		
Non-EU country	82.7	(66.8-98.6)	81.0	(65.0-97.0)		
Region						
Flanders	78.0	(73.4-82.6)	76.1	(71.2-80.9)	71.2	(70.6-71.8)
Brussels	75.8	(68.4-83.2)	66.7	(58.5-74.9)	59.3	(57.40-61.2)
Wallonia	70.3	(64.9-75.6)	57.3	(51.4-63.2)	61.8	(61.0-62.8)
Income					*N.A*	
Low	71.9	(66.6-77.1)	66.2	(60.7-71.6)		
High	79.4	(74.5-84.3)	74.6	(69.2-80.0)		
Health status					*N.A*	
Good to very good	79.7	(75.8-83.6)	74.0	(69.7-78.2)		
Very bad to fair	65.0	(58.6-71.3)	58.6	(51.8-65.4)		

BHIS = Belgian Health Interview Survey; BCHI = Belgian Compulsory Health Insurance
EPS = Permanent Sample (random sample of the Belgian Compulsory Health Insurance data); N.A = Not available.

Table 4.1.2 summarizes both the selection and reporting biases. A significant difference between the BHIS-BCHI and the EPS mammography uptake reimbursement rates is observed overall. The absolute and relative size of the selection bias is 5.72 percentage points (95% CI: 2.06-9.38) and 8.93% (95%CI: 3.21- 14.64), respectively. No significant differences were detected between subgroups.

Also for the reporting bias, a significant difference between self-reported and reimbursement information in the BHIS-BCHI is observed. The absolute size is 5.74 percentage points (95% CI: 0.75-10.7) and the relative size is 8.22% (95% CI: 0.76- 15.68), respectively. A subgroup analyses indicates that the mammography uptake is over reported by 14% for women aged 50-59 years, 11% for those highly educated, 8% for women born in Belgium and 23% for those residing in Wallonia. This over-reporting is confirmed by the RRR in the related subgroups.

Table 4.1.2: Estimated bias in mammography uptake in the last 2 years among women aged 50-69 years in the BHIS-BCHI linked sample, HISlink 2013, Belgium

	Estimated bias[a]				
	Selection bias[b] (%)		Reporting bias[c] (%)		
	Absolute[d] (95% CI)	Relative[e] (95% CI)	Absolute[f] (95% CI)	Relative[g] (95% CI)	RRR (95% CI)
Overall	5.72 (2.06-9.38)*	8.93 (3.21-14.64)*	5.74 (0.75-10.72)*	8.22 (0.76-15.68)*	1.08 (1.02-1.15)*
Age (years)					
50-59	1.56 (-3.50-6.62)	2.32 (-5.23-9.87)	9.42 (2.67-16.17)*	13.72 (3.12-24.33)*	1.14 (1.05-1.23)*
60-69	4.04 (-1.27-9.34)	6.02 (-1.91-13.95)	1.67 (-0.57-9.06)	2.35 (-8.16-12.87)	1.02 (0.94-1.11)
Educational level	N.A[h]				
Low			5.22 (-5.09-15.53)	8.56 (-9.11-26.23)	1.09 (0.95-1.24)
Middle			2.99 (-5.67-11.65)	4.08 (-7.97-16.12)	1.04 (0.95-1.14)
High			7.98 (0.62-15.34)*	10.83 (0.22-21.45)*	1.11 (1.02-1.21)*
Place of birth	N.A[h]				
Belgium			5.86 (0.63-11.10)*	8.36 (0.56-16.17)*	1.08 (1.02-1.15)*
EU country			5.67 (-14.66-26.01)	9.86 (-27.25-46.97)	1.10 (0.90-1.34)
Non-EU country			1.70 (-19.67-23.07)	2.10 (-24.54-28.74)	1.02 (0.83-1.25)
Region	N.A[h]				
Flanders	4.86 (-0.03-9.75)	6.82 (-0.06-13.70)	1.95 (-4.75-8.66)	2.57 (-6.36-11.50)	1.03 (0.95-1.11)
Brussels	7.43 (-0.89-15.75)	12.53 (-1.60-26.67)	9.12 (-1.83-20.01)	13.67 (-4.00-31.34)	1.14 (1.00-1.29)
Wallonia	-4.56 (-10.51-1.38)	-7.38 (-16.98-2.22)	12.95 (5.01-20.89)*	22.60 (6.95-38.25)*	1.23 (1.10-1.36)*
Income	N.A[h]				
Low			5.69 (-1.86-13.25)	8.61 (-3.33-20.54)	1.09 (0.99-1.19)
High			4.82 (-2.54-12.19)	6.46 (-3.76-16.69)	1.05 (0.97-1.13)
Health status	N.A[h]				
Good to very good			5.69 (-0.07-11.44)	7.69 (-0.41-15.79)	1.08 (1.01-1.15)
Very bad to fair			6.41 (-3.15-15.98)	10.96 (-6.32-28.23)	1.11 (0.98-1.26)

[a] *Computed before rounding the percentages.*
[b] *Computed by comparing the percentage of women with a mammography reimbursement in the BHIS-BCHI linked sample and in the EPS data.*
[c] *Computed by comparing the percentage of self-reported mammography uptake and mammography reimbursement in the BHIS-BCHI linked data.*

[d] *Absolute difference in the prevalence of mammography reimbursement rates in the BHIS-BCHI linked sample and similar estimates from the EPS data.*
[e] *Relative excess in percentage , computed as the differences between the percentage of women with a mammography reimbursement in the BHIS-BCHI linked sample and in the EPS data, divided by the percentage from the EPS data .*
[f] *Absolute difference in the prevalence of self-reported mammography uptake and reimbursement rate in the BHIS-BCHI linked data.*
[g] *Relative excess in percentage, computed as the difference between the percentage of self-reported mammography uptake and mammography reimbursement in the BHIS-BCHI linked data, divided by the percentage of reimbursement in the BHIS-BCHI linked data*
[h] *N.A = Not available.*
**Significant result (p<0.05).*

Table 4.1.3 reports the more common measures of agreement related to the reporting bias. The sensitivity was excellent overall (93.7%) and across subgroups except for women born in other EU countries and for those reporting a poor perceived health. Whereas the specificity was poor (66.4%) overall and did not exceed 70% in most of the subgroups. When the time frame was moved from 2 to 3 years, the specificity increased to 83%. The overall agreement was 84% (result not shown) and the kappa statistics was 0.63. The PPV was good overall and in all subgroups except for women born in non-EU countries where it was excellent. The NPV was above 80% in all subgroups but fair for women aged 60-69 years and those with middle educational level.

Table 4.1.3: Measures of validity of self-reported mammography uptake using administrative data as gold standard (BHIS-BCHI linked), HISlink 2013, Belgium

Characteristics	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	Kappa statistic
Overall	93.7 (91.5-95.9)	66.4 (59.7-73.1)	86.6 (83.5-89.6)	82.0 (76.2-87.8)	0.63 (0.58-0.68)
Age (years)					
50-59	95.9 (93.1-98.6)	61.0 (51.5-70.4)	84.3 (79.9-88.7)	87.1 (78.9-95.3)	0.63 (0.56-0.70)
60-69	91.4 (87.8-94.9)	73.0 (64.0-81.9)	89.3 (85.3-93.3)	77.4 (68.9-85.9)	0.64 (0.56-0.71)
Education					
Low	98.4 (84.1-94.7)	70.0 (58.9-81.1)	82.3 (75.2-89.5)	80.8 (71.1-90.6)	0.60 (0.50-0.70)
Middle	92.1 (87.7-96.4)	66.9 (54.1-79.7)	88.5 (83.6-93.3)	75.4 (62.1-88.7)	0.59 (0.50-0.69)
High	97.5 (95.0-99.9)	62.6 (50.7-74.5)	87.9 (83.3-92.6)	89.8 (80.7-98.9)	0.70 (0.62-0.78)
Place of birth					
Belgium	93.8 (91.4-96.1)	65.7 (58.6-72.8)	86.5 (83.3-89.7)	81.8 (75.6-88.0)	0.64 (0.58-0.69)
EU countries	88.0 (78.1-97.8)	70.4 (50.3-90.5)	80.1 (66.0-94.2)	81.2 (64.6-97.9)	0.53 (0.34-0.72)
Non-EU country	99.0 (96.9-100)	86.7 (67.9-100)	96.9 (92.9-100)	95.2 (83.8-100)	0.77 (0.55-0.98)
Region					
Flanders	92.6 (89.6-95.7)	68.4 (57.5-79.3)	90.3 (86.5-94.1)	74.5 (64.7-84.2)	0.61 (0.52-0.70)
Brussels	98.5 (96.7-100)	69.6 (55.7-83.5)	86.7 (80.0-93.3)	95.9 (90.9-100)	0.72 (0.62-0.83)
Wallonia	95.5 (92.8-98.1)	63.6 (54.8-72.3)	77.9 (71.9-83.8)	91.3 (86.2-96.4)	0.60 (0.53-0.67)
Income category					
Low	93.7 (90.3-97.1)	70.9 (62.7-79.0)	86.3 (82.0-90.5)	85.2 (77.5-92.9)	0.64 (0.57-0.71)
High	93.6 (90.3-96.9)	62.2 (49.5-74.9)	87.9 (83.2-92.6)	76.7 (67.4-86.0)	0.61 (0.53-0070)
Health status					
Good to very good	95.2 (92.9-97.5)	64.6 (55.7-73.4)	88.4 (85.1-91.8)	82.6 (74.8-90.4)	0.65 (0.58-0.71)
Very bad to fair	89.1 (83.7-94.5)	69.1 (58.9-79.3)	80.3 (73.3-87.3)	81.8 (72.5-91.0)	0.60 (0.51-0.69)

The results of the multivariate logistic are shown in Table 4.1.4. Inaccurate self-reported mammography uptake is more common among women born in a non-EU country (OR = 2.81, 95% CI: 1.54-5.13), people with a high household income (OR= 1.27, 95% CI:1.02-1.60) and those reporting very bad to fair perceived health.

Table 4.1.4: Adjusted odds ratios (with 95% CI) of inaccurate self-reported mammography uptake in the past 2 years (defined as over-reporting or under-reporting). Results of multivariate logistic regression, HISlink 2013, Belgium

Characteristics	OR (95% CI)
Age (years)	
50-59	1.00
60-69	1.05 (0.86-1.29)
Educational level	
Low	1.32 (0.97-1.80)
Middle	1.06 (0.79-1.41)
High	1.00
Place of birth	
Belgium	1.00
EU country	1.35 (0.77-2.35)
Non-EU country	2.81 (1.54-5.13)*
Region	
Flanders	1.00
Brussels	1.29 (0.92-1.82)
Wallonia	1.08 (0.83-1.41)
Income	
Low	1.00
High	1.27 (1.02-1.60)*
Health status	
Good to very good	1.00
Very bad to fair	1.41 (1.14-1.73)*

**Significant result (p<0.05).*

4.1.5. Discussion

The main objective of this study was to assess the validity in terms of selection and reporting bias of self-reported mammography uptake in the BHIS. In the BHIS as in other interview surveys, the validity of self-reported information depends both on the selection and reporting bias. Our results indicate that the mammography uptake in the BHIS is significantly affected by both types of biases. Therefore, cautiousness is needed when using self-reported estimates as the sole method to quantify mammography coverage.

Due to the compulsory nature of the Belgian health insurance and the fact that the Belgian federal and regional governments signed a protocol agreement in 2001 for an organized screening program for women aged 50-69 years, to be organized by the regional government with appropriate financial resources supplied by the federal government, it can be stated that indicators based on the BCHI are quite reliable.

We found a significant selection bias. The relative overestimation of self-reported information was 9% overall.

Mammography uptake is also significantly affected by reporting bias in the same direction and in a comparable manner. Indeed, the relative overestimation of the percentage from the BHIS is 8% overall. This significant overestimation is observed across subgroups. Theoretically, the over-reporting could be partially due to an incomplete recording in the BCHI [5;39], but this is highly unlikely because for the financial management of the health insurance accurate data are essential. Therefore, administrative mistakes made by health insurance employees can be considered to be negligible. Another potential explanation is the underestimation of the timeframe since the last exam. This phenomenon, also called “telescoping” (i.e., remembering that an event occurred more recently than it actually did), is the most consistent finding among studies comparing self-reports with medical or administrative data sources [12;20;40].

The poor specificity found in our study (<70%) suggesting a higher rate of false positives could confirm the hypothesis of telescopic bias. We found that the telescopic bias represents almost half of the false positive cases. Indeed, if the time frame was moved from 2 to 3-years, the specificity would have been 83%. Over-reporting may also occur because adhesion to screening recommendations is perceived to be

socially desirable [12]. As opposed to findings in the literature [6;13], our results did not show that over-reporting mammography uptake occurred more often among women with a lower socio economic status. On the contrary, our results suggested that women with high household income level are more likely to inaccurately report (over-report) their mammography uptake.

When adjusted for other variables, women born in a non-EU country are more likely to inaccurate report (over-report) their mammography uptake as opposed of results from Tables 4.1.2 and 4.1.3.

In the complete case analysis (results not shown) , only the place of birth was significantly associated with inaccurate report of mammography uptake, probably because of loss of power due to drop out of missing values. Although the other variables were not significantly associated with the outcome, the direction of the effect remain unchanged as in analysis after multiple imputation.

Other validation studies have found results that are in line with those in our study. In their meta-analysis, Howard et al. [12] estimated the pooled sensitivity and the pooled specificity to 95% and 62% respectively. In another meta-analysis, Anderson et al. [5] also found excellent sensitivity (96%) but moderate specificity (61%). In another study, the specificity was much lower (45%) while the sensitivity was comparable [40]. The authors explained this difference by the higher underestimation of the time elapsed since the last exam.

An important advantage of our study compared to most other studies is the fact that it was conducted in a representative sample of the population. The most common data used as gold standard in validation studies are medical records [12;32;40], which can be considered as more accurate than administrative data. However, medical data could be too difficult and expensive to obtain for population estimates. In our context, the use of administrative data as the gold standard is acceptable since they give exhaustive and accurate information on the number of mammograms that are carried out. Therefore, similar measures of validity (sensitivity, specificity) can be used as in studies that used medical records data as gold standard. The overall agreement (84.4% - result not shown) and the kappa statistic (0.63) as measures of reliability observed in our study were comparable to those in other studies [32;40].

Another important strength of the current study is that we assessed concomitantly the selection and the reporting bias.

Some limitations of this study need to be highlighted. First, no distinction could be made between mammograms as part of a screening program and opportunistic mammograms in the BHIS. Moreover, because opportunistic screening mammograms are often miscoded as diagnostic mammograms for reimbursement purposes in the BCHI, we were unable to distinguish screening mammograms from diagnostic mammograms. However, since the proportion of diagnostic mammograms among all mammograms is quite low, the rate of mammograms outside the screening is an acceptable proxy of the opportunistic screening. So, the actual indicator that was assessed was "having had mammograms", including both screening and opportunistic mammograms. The share of each type has never been measured in Belgium. In this study, we assumed that the largest part of the mammograms undergone between 50 and 69 is made for screening purposes, and therefore we used this information as a proxy of the breast cancer screening. Second, only a subpopulation of the BHIS participants (women aged 50 to 69 years) is analyzed. Ideally, a re-calibration of sample weights will be optimal. Unfortunately, because of the limited number of demographic variables in the reference dataset, this was not possible. Third, although it may seem more logical if we would have compared estimates obtained in the BHIS with screening information from the complete population, the data protection authority does not allow the use of exhaustive information from the BCHI if equally reliable information can be obtained from the EPS. As the EPS is a large sample and selected through a random procedure, it can be assumed that the EPS estimates perfectly match the indicators that would have been obtained from the total population.

This study has implications for public health policy-makers. Self-reported mammography uptake is not the most accurate method to track the national screening coverage rate and to determine the adherence to the national or international guidelines or attainment of goals. Therefore, the self-reported mammography uptake should be interpreted with caution and when possible objective data should be used.

Despite the moderate validity of mammography uptake in the BHIS , this data source still has an added value since it provides information on the sociodemographic determinants of the mammography attendance, and the link with health behaviors and other health outcomes.

4.1.6. Conclusions

In the BHIS as in other interview surveys, the validity of self-reported information depends both on the selection and reporting bias. Our results indicate that the mammography uptake in the BHIS is significantly affected by both types of biases. Therefore, cautiousness is needed when using self-reported estimates as the sole method to quantify mammography coverage. Despite the moderate validity of mammography uptake in the BHIS, this data source still has an added value since it provides information on the sociodemographic determinants of the mammography attendance, and the link with health behaviors and other health outcomes. Further dedicated studies are needed to confirm our findings.

Key points

- Mammography uptake is overestimated in the Belgian health interview survey
- Although the sensitivity of self-reported information of mammography uptake is excellent, the fair specificity indicates a higher rate of false positive, especially in some subgroups
- Despite their moderate validity, data from the Belgian health interview survey are still useful to identify the determinants of breast cancer screening and to monitor health inequalities over time in this field
- Public health policy-makers should consider both data sources when assessing mammography uptake: administrative data to monitor overall changes and geographic differences; survey data to better understand differential in uptake.

4.1.7. Bibliography

1. Ferlay J, Soerjomataram I, Dikshit R et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer 2015;136(5):E359-E386.
2. International Agency for Research on Cancer (IARC). Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. Press Release N° 263 12/09/2018, in press.
3. Torre LA, Siegel RL, Ward EM, Jemal A. Global cancer incidence and mortality rates and trends-An update. Cancer Epidemiology and Prevention Biomarkers 2016;25(1):16-27.
4. Martín-Sánchez JC, Lunet N, González-Marrón A et al. Projections in breast and lung cancer mortality among women: A Bayesian analysis of 52 countries worldwide. Cancer research 2018;78(15):4436-42.
5. Anderson, Johanna, Bourne, Donald, Peterson, Kim, and Mackey, Katherine. Evidence Brief: Accuracy of Self-report for Cervical and Breast Cancer Screening. VA ESP Project #09-199. 2019.
6. Lofters A, Vahabi M, Glazier RH. The validity of self-reported cancer screening history and the role of social disadvantage in Ontario, Canada. BMC Public Health 2015;15(1):28.
7. Hanley JA, Hannigan A, O'Brien KM. Mortality reductions due to mammography screening: Contemporary population-based data. PLoS One 2017;12(12):e0188947.
8. Njor S, Nyström L, Moss S et al. Breast cancer mortality in mammographic screening in Europe: a review of incidence-based mortality studies. Journal of medical screening 2012;19(1_suppl):33-41.
9. Perry N, Puthaar E. European guidelines for quality assurance in breast cancer screening and diagnosis.European Communities, 2006.
10. Autier P, Boniol M, Gavin A, Vatten LJ. Breast cancer mortality in neighbouring European countries with different levels of screening but similar access to treatment: trend analysis of WHO mortality database. Bmj 2011;343:d4411.
11. Council of the European Union. Council Recommendation of 2 December 2003 on cancer screening (2003/878/EC). Off J Eur Union 2003;327:34-38.
12. Howard M, Agarwal G, Lytwyn A. Accuracy of self-reports of Pap and mammography screening compared to medical record: a meta-analysis. Cancer Causes & Control 2009;20(1):1.
13. Lofters AK, Moineddin R, Hwang SW, Glazier RH. Does social disadvantage affect the validity of self-report for cervical cancer screening? International journal of women's health 2013;5:29.

14. Puddu M, Demarest S, Tafforeau J. Does a national screening programme reduce socioeconomic inequalities in mammography use? International journal of public health 2009;54(2):61-68.
15. Commission of the European Communities. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the regions on Action Against Cancer: European Partnership. Brussels, COM(2009) 291/4.22. 2009.
16. Dimitrova, Nadya, Parkinson, Zuleika Saz, Bramesfeld, Anke, Ulutürk, Asli, Bocchi, Giulia, Lòpez-Alcalde, Jesús, Pylkkanen, Liisa, Neamtiu, Luciana, Ambrosio, Massimo, and Deandrea, Silvia. The European Breast Guidelines. 2016.
17. Office of Disease Prevention and Health Promotion (ODPHP). Healthy People 2020. 12-10-2018. Available at: https://www.healthypeople.gov/2020/topics-objectives/topic/cancer/objectives (15 July 2020, date last accessed).
18. Sabatino SA et al. CDC: Cancer screening rates remain below Healthy People 2020 targets. MMWR. 2015;64:464-68.
19. Champion VL, Menon U, McQuillen DH, Scott C. Validity of self-reported mammography in low-income African-American women. American Journal of Preventive Medicine 1998;14(2):111-17.
20. Cronin KA, Miglioretti DL, Krapcho M et al. Bias associated with self-report of prior screening mammography. Cancer Epidemiology and Prevention Biomarkers 2009;18(6):1699-705.
21. Van Hal G, Thibaut A, Matthyssen M, Weyler J. Linking a breast cancer screening data base with a cancer registry in Antwerp, Belgium. Archives of Public Health 2000;58:307-19.
22. Belgian Cancer Registry Brussels. Cancer Incidence in Belgium, 2008. 2011.
23. Renard F, Tafforeau J, Deboosere P. Premature mortality in Belgium in 1993-2009: leading causes, regional disparities and 15 years change. Archives of Public Health 2014;72(1):34.
24. Vrijens F, Renard F, Camberlin C et al. Performance of the Belgian health system-report 2015. Brussels: supplement health Services Research (HSR) 2016.
25. InterMutualistic Agency (IMA). IMA Atlas. 2018. Available at: http://atlas.aim-ima.be/base-de-donnees (15 July 2020, date last accessed).
26. Tafforeau, Jean. Enquête de santé par interview, Belgique 2008. Le dépistage du cancer. 2008.
27. Van der Heyden J, Charafeddine R, De Bacquer D, Tafforeau J, Van Herck K. Regional differences in the validity of self-reported use of health care in Belgium: selection versus reporting bias. BMC medical research methodology 2016;16(1):98.

28. Demarest S, Van der Heyden J, Charafeddine R, Drieskens S, Gisle L, Tafforeau J. Methodological basics and evolution of the Belgian health interview survey 1997-2008. Archives of Public Health 2013;71(1):24.
29. Van der Heyden J. Validity of the assessment of population health and use of health care in a national health interview survey. 2017.
30. Rupp I, Triemstra M, Boshuizen HC, Jacobi CE, Dinant HJ, van den Bos GA. Selection bias due to non-response in a health survey among patients with rheumatoid arthritis. The European Journal of Public Health 2002;12(2):131-35.
31. Moss CA. Colorectal cancer screening in the Iowa Research Network (IRENE): a validity assessment of patient self-report of up-to-date status. 2014.
32. Tiro JA, Sanders JM, Shay LA et al. Validation of self-reported post-treatment mammography surveillance among breast cancer survivors by electronic medical record extraction method. Breast cancer research and treatment 2015;151(2):427-34.
33. Reiter PL, Katz ML, Oliveri JM, Young GS, Llanos AA, Paskett ED. Validation of self-reported colorectal cancer screening behaviors among Appalachian residents. Public Health Nursing 2013;30(4):312-22.
34. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37(5):360-363.
35. McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica 2012;22(3):276-82.
36. Charafeddine R, Demarest S, Cleemput I, Van Oyen H, Devleesschauwer B. Gender and educational differences in the association between smoking and health-related quality of life in Belgium. Preventive medicine 2017;105:280-286.
37. Althouse AD. Adjust for multiple comparisons? It's not that simple. The Annals of thoracic surgery 2016;101(5):1644-5.
38. Volken T. Second-stage non-response in the Swiss health survey: determinants and bias in outcomes. BMC Public Health 2013;13(1):167.
39. Ferrante JM, Ohman-Strickland P, Hahn KA et al. Self-report versus medical records for assessing cancer-preventive services delivery. Cancer Epidemiology and Prevention Biomarkers 2008;17(11):2987-94.
40. Caplan LS, McQueen DV, Qualters JR, Leff M, Garrett C, Calonge N. Validity of women's self-reports of cancer screening test utilization in a managed care population. Cancer Epidemiology and Prevention Biomarkers 2003;12(11):1182- 87.

4.2. COMPARING ADMINISTRATIVE AND SURVEY DATA FOR ASCERTAINING CHRONIC DISEASE PREVALENCE

The findings of this paper were published as:

Berete F, Demarest S, Charafeddine R, Bruyère O and Van der Heyden J. Comparing health insurance data and health interview survey data for ascertaining chronic disease prevalence in Belgium. Archives of Public Health 78.1 (2020): 1-9.

Berete *et al. Archives of Public Health* (2020) 78:120
https://doi.org/10.1186/s13690-020-00500-4

Archives of Public Health

RESEARCH Open Access

Comparing health insurance data and health interview survey data for ascertaining chronic disease prevalence in Belgium

Finaba Berete[1,2*], Stefaan Demarest[1], Rana Charafeddine[1], Olivier Bruyère[3] and Johan Van der Heyden[1]

Abstract

Background: Health administrative data were increasingly used for chronic diseases (CDs) surveillance purposes. This cross sectional study explored the agreement between Belgian compulsory health insurance (BCHI) data and Belgian health interview survey (BHIS) data for asserting CDs.

Methods: Individual BHIS 2013 data were linked with BCHI data using the unique national register number. The study population included all participants of the BHIS 2013 aged 15 years and older. Linkage was possible for 93% of BHIS-participants, resulting in a study sample of 8474 individuals. For seven CDs disease status was available both through self-reported information from the BHIS and algorithms based on ATC-codes of disease-specific medication, developed on demand of the National Institute for Health and Disability Insurance (NIHDI). CD prevalence rates from both data sources were compared. Agreement was measured using sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) assuming BHIS data as gold standard. Kappa statistic was also calculated. Participants' sociodemographic and health status characteristics associated with agreement were tested using logistic regression for each CD.

Results: Prevalence from BCHI data was significantly higher for CVDs but significantly lower for COPD and asthma. No significant difference was found between the two data sources for the remaining CDs. Sensitivity was 83% for CVDs, 78% for diabetes and ranged from 27 to 67% for the other CDs. Specificity was excellent for all CDs (above 98%) except for CVDs. The highest PPV was found for Parkinson's disease (83%) and ranged from 41 to 75% for the remaining CDs. Irrespective of the CDs, the NPV was excellent. Kappa statistic was good for diabetes, CVDs, Parkinson's disease and thyroid disorders, moderate for epilepsy and fair for COPD and asthma. Agreement between BHIS and BCHI data is affected by individual sociodemographic characteristics and health status, although these effects varied across CDs.

Conclusions: NHIDI's CDs case definitions are an acceptable alternative to identify cases of diabetes, CVDs, Parkinson's disease and thyroid disorders but yield in a significant underestimated number of patients suffering from asthma and COPD. Further research is needed to refine the definitions of CDs from administrative data.

Keywords: Chronic diseases, Health administrative data, Data linkage, Validity, HEALTH insurance data, Chronic diseases ascertainment

4.2.1. Abstract

Background

Health administrative data were increasingly used for chronic diseases (CDs) surveillance purposes. This cross sectional study explored the agreement between Belgian compulsory health insurance (BCHI) data and Belgian health interview survey (BHIS) data for asserting CDs.

Methods

Individual BHIS 2013 data were linked with BCHI data using the unique national register number. The study population included all participants of the BHIS 2013 aged 15 years and older. Linkage was possible for 93% of BHIS-participants, resulting in a study sample of 8474 individuals. For seven CDs disease status was available both through self-reported information from the BHIS and algorithms based on ATC-codes of disease-specific medication, developed on demand of the National Institute for Health and Disability Insurance (NIHDI). CD prevalence rates from both data sources were compared. Agreement was measured using sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) assuming BHIS data as gold standard. Kappa statistic was also calculated. Participants' sociodemographic and health status characteristics associated with agreement were tested using logistic regression for each CD.

Results

Prevalence from BCHI data was significantly higher for CVDs but significantly lower for COPD and asthma. No significant difference was found between the two data sources for the remaining CDs. Sensitivity was 83% for CVDs, 78% for diabetes and ranged from 27% to 67% for the other CDs. Specificity was excellent for all CDs (above 98%) except for CVDs. The highest PPV was found for Parkinson's disease (83%) and ranged from 41% to 75% for the remaining CDs. Irrespective of the CDs, the NPV was excellent. Kappa statistic was good for diabetes, CVDs, Parkinson's disease and thyroid disorders, moderate for epilepsy and fair for COPD and asthma. Agreement between BHIS and BCHI data is affected by individual sociodemographic characteristics and health status, although these effects varied across CDs.

Conclusions

NHIDI's CDs case definitions are an acceptable alternative to identify cases of diabetes, CVDs, Parkinson's disease and thyroid disorders but yield in a significant underestimated number of patients suffering from asthma and COPD. Further research is needed to refine the definitions of CDs from administrative data.

Keys words: Chronic diseases, health administrative data, data linkage, validity, health insurance data, Chronic diseases ascertainment.

4.2.2. Background

Chronic diseases (CDs) represent an important concern for public health policy. Indeed, their prevalence is constantly increasing and they are by far the leading cause of mortality in Europe, representing 77% of the total disease burden and 86% of all deaths [1].

An important prerequisite for the CDs management is to be able to identify, in a valid, simple and inexpensive way, the population with CDs that need proactive and planned care [2]. For this purpose, population-based data for routine monitoring of CDs prevalence are fundamental to describe the burden of disease and to plan and evaluate disease prevention, treatment and management strategies and by defining target populations [3,4].

Prevalence of CDs is often estimated using population health surveys, disease registers, hospitalization or outpatient records [3–8]. Besides these traditional methods, health administrative databases have been used as an alternative, efficient source of data for CDs surveillance [4,5,9,10]. Health administrative databases can be accessed easily and quickly, associated costs are low and they are quite exhaustive. In some cases such databases can be used to provide cross-sectional and longitudinal data on the prevalence and incidence of diseases in the entire population [10]. The use of such data is very challenging [11] yet due to the opportunity they provide, they have often been used for surveillance purposes. For instance, in France, the French national health insurance information system (Système National de Données de Santé – SNDS) has been used to develop the Diabetes National Surveillance System which serves as a base to estimate the national prevalence of pharmacologically treated diabetes and the incidence of

diabetes-related complications, as well as their temporal trends and their territorial variations [12]. To estimate these indicators, a diabetes case definition algorithm based on antidiabetic drug consumption was applied [4]. Drug use data, especially prescription drugs, have also been frequently used to estimate CDs prevalence [5,7,13].

In Belgium, the prevalence of specific CDs is usually assessed, based on data gathered through the Belgian health interview survey (BHIS), conducted every 5-years. Next to this, other sources such as hospital discharge data, disease-specific registries (e.g., Belgian cancer registry), sentinel practice networks (e.g., Intego sentinel GP network), also represent important tools to obtain prevalence/incidence rates of CDs.

More than 99% of the Belgian population is covered by the Belgian compulsory health insurance (BCHI). The BCHI database provides detailed and complete information on the reimbursement of health care costs for almost the entire population. Such information is widely used by important actors in the health field, such as the National Institute for Health and Disability Insurance (NIHDI), the Belgian health care knowledge centre and the Federal planning bureau tor studying and planning topics mainly related to health care costs and expenditures. Although these data are not meant for epidemiological purposes, BCHI data are also used to estimate the prevalence of some CDs at population level [14].

At the initiative of the NIHDI, a panel of experts (mainly clinicians) have developed algorithms based on prescribed medication dispensed in public pharmacies to construct indicators of CDs. The algorithms are all based on a minimum consumption of 90 DDD (Defined Daily Dose) during one calendar year of drugs of certain (sub)classes of ATC (Anatomical Therapeutic Chemical), often in combination with the minimum age of the patient [15].

These indicators of CDs are useful for the NIHDI, to identify specific patient populations. However, since their development, they have only been validated qualitatively. To our knowledge, only one study has compared the prevalence of diabetes mellitus and thyroid disorders from BHIS, BCHI and diagnostic codes in Flanders [6].

The main objective of this study was to assess agreement between health administrative and self-reported cases definitions of diabetes, asthma, chronic obstructive pulmonary disease (COPD), cardiovascular diseases including hypertension (CVDs), Parkinson's disease, thyroid disorders and epilepsy in the Belgian population, assuming self-reported data as a gold standard. The aforementioned CDs were chosen because they are common diseases with a lower risk of misreporting by BHIS participants and because they are generally treated with specific drugs which are more or less specific for the disease. Furthermore, we also sought to determine the subject sociodemographic and health status characteristics that may affect the agreement between the two data sources.

4.2.3. Methods

Study design and population

This is a descriptive cross-sectional study. The study population included all participants of the Belgian health interview survey (BHIS) 2013 aged 15 years and older (n=9112).

Data sources

Date were derived from the HISLINK 2013 study, an individual linkage between the Belgian health interview survey (BHIS) 2013 data and the Belgian compulsory health insurance data (BCHI) from 2012 to 2018.

The BHIS is a national, cross-sectional household survey conducted every 5 years since 1997 by Sciensano, the Belgian health institute, among a representative sample of Belgian residents. Participants are selected from the national population register through a multistage stratified sampling procedure. The participation rate in the survey was 57% at the household level. In the BHIS, information is collected on health status, health behavior, health care consumption, sociodemographic characteristics and use of medicines. The detailed methodology of the survey is described elsewhere [16].

The BCHI data contain exhaustive and detailed information on the reimbursed health expenses of over 99% of the total population. The database also includes a limited amount of socio-demographic information. The BCHI data were provided by the Intermutualistic Agency (IMA). IMA is a joint venture of the seven national sickness funds and collects and manages all data on healthcare expenditures as well as

prescription information on reimbursed medicines (Pharmanet data) [17]. Pharmanet logs all data on reimbursed dispensed medication from public pharmacies in Belgium. Pharmanet data include information on the date of dispensing, the quantity per package (QPP), the daily defined dose (DDD) and the national code number of the medicine (CNK codes) which allows to link each medicine to its ATC-code. The list of ATC codes per CNK codes was provided by the NIHDI.

Individual BHIS 2013 data were linked with BCHI data using the unique national register number. The study population included all participants of the BHIS 2013 aged 15 years and older (n=9112). The linkage was possible for 93% of them, resulting in a final sample of 8474 individuals. The HISLINK 2013 was used because it was the most recent linked database available at the moment of this study.

Identification of chronic diseases

The prevalence information from BHIS was collected using a list of CDs (35 in total) based on the following question: *"Have you suffered during the last 12 months from the following disease?".* Since there is no specific indicator for CVDs in the BHIS, we considered a person to have CVDs (including hypertension) when they reported having had in the past 12 months at least one of the following CDs: myocardial infarction, coronary disease, hypertension, stroke, or other serious heart diseases.

In the BCHI data, the NIHDI algorithms were used to ascertained cases of CDs. In these algorithms, CDs cases were identified based on the ATC-codes of dispensed medication in public pharmacies, using the WHO guidelines on the ATC classification system [18]. So, a CD was assigned to a participant if the total of DDDs reimbursed for all selected ATC-codes used in the treatment for this CD is greater or equal to 90 [15] in the past 12 months preceding the participation in the BHIS. The selected ATC-codes for each CD are presented in Table 4.2.1.

Statistical analyses

We calculated the weighted prevalence rates from both data sources for the 7 selected CDs. The delta method [19] was applied to test if there was a significant difference between the estimates of both sources.

The agreement was measured by estimating sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) and their 95% CI, assuming BHIS data as gold standard. Sensitivity was defined as the percentage of true positive cases an algorithm detects among all positive disease cases. Positive disease cases are BHIS respondents who reported having the specified disease. Specificity was defined as the percentage of true negative cases an algorithm detects among all the negative disease cases. Negative disease cases are BHIS respondents who did not report having the specified disease. Positive predictive values (PPVs) and negative predictive values (NPVs) are also reported for each chronic disease algorithm. PPV refers to the percentage of individuals with a positive result for an algorithm among those who reported having the disease. NPV refers to the percentage of individuals with a negative result for an algorithm who did not report having the disease [20].

Furthermore, Kappa values were calculated to differentiate between true agreement and agreement produced by chance. Kappa values were interpreted as follows: $\kappa \leq 0.40$, fair-to-poor agreement; $\kappa = 0.41$ to 0.60, moderate agreement; $\kappa = 0.61$ to 0.80, substantial agreement; and $\kappa = 0.81$ to 1.00, almost perfect agreement [21].

Sensitivity analyses were conducted by repeated analyses for different cut-off points of the DDD for all the CDs.

Finally, univariable and multivariable logistic regression analysis were performed for each CD (except for the Parkinson's disease because of small number of cases unable to provide reliable estimates) to further investigate the effect of respondent's characteristics on the total agreement (true positive or true negative) between BHIS and BCHI data sources. Participants characteristics included in the model are: gender, age-group (15-34, 35-54, 55-74 and 75+ years), education (low, intermediate, high), nationality (Belgian, EU-countries, other countries), household income (quintile), region of residence (Flanders, Brussels, Wallonia), self-perceived health (good to very good, very bad to fair), presence of multimorbidity (yes/no) and polypharmacy defined as simultaneous use 5 medicines or more on a typical day (yes/no).

A two-sided alpha level of 0.05 was considered statistically significant. All analyses were performed using SAS 9.4 (SAS Institute Inc., Cary, NC, USA) and Stata 16.1 and taking into account the survey settings.

Ethics statement

As mentioned above, this study was carried out using the individual linkage between the BHIS 2013 data and the BCHI data. The BHIS 2013 was carried out in line with the Belgian privacy legislation and has been approved by the ethics committee of the University hospital of Ghent on October, 1st 2012 (advice EC UZG 2012/658). The participation to BHIS is voluntary. There was no formal written and signed consent foreseen as participation was considered as consent. In addition, for the data linkage, an authorization was obtained from the Information Security Committee (local reference: Deliberation No. 17/119 of December 19, 2017, amended on September 3, 2019).

This study is reported according to the STROBE statement.

4.2.4. Results

Table 4.2.1 summarizes the CDs with identification questions in the BHIS data source and the assigned ATC-codes in the BCHI data source.

Table 4.2.1: Survey questions and ATC prescription drug codes for chronic disease case ascertainment, HISlink 2013, Belgium

Chronic diseases	Survey questions : " Have you suffered during the last 12 months from…"	ATC-codes
Diabetes mellitus	Diabetes?	A10A A10B
Cardiovascular diseases	Myocardial infarction? Coronary disease? Hypertension? Stroke? Other serious heart disease?	C01 C02 C03 C07 C08 C09
COPD	COPD?	R03BB R03DA04 R03A[a] R03BA[a]
Asthma	Asthma?	R03DC01 R03DC03 R03DX05 R03A[b] R03BA[b]
Parkinson's disease	Parkinson's disease?	N04AB N04AC N04B
Epilepsy	epilepsy?	N03
Thyroid disorders	thyroid disorders?	H03AA

[a] *For people aged <= 50 years;* [b] *For people aged > 50 years*

Characteristics of the study population, unweighted and weighted to reflect the general Belgian population in terms of age, gender and region are presented in Table A1 (supplementary material). More than half of the population perceived their health to be good to very good, 15% suffers from multimorbidity and one person out of ten simultaneous uses 5 medicines or more on a one day reference period.

Table 4.2.2 shows the prevalence of CDs in the population by data source. The prevalence rates obtained from administrative data source were significantly higher than those obtained from survey data for CVDs (including hypertension), but on the contrary, they were significantly lower for COPD and asthma. No significant difference was found between the two data sources for the remaining CDs.

Table 4.2.2: Prevalence (weighted percentages) of chronic diseases among the population aged 15 years and over by data source, HISlink 2013, Belgium

Chronic disease	Prevalence in BHIS (E1)	Prevalence in BCHI (E2)	Absolute difference[a] (E1-E2)	Relative difference[a] (E1-E2)/E2
	% (95% CI)	% (95% CI)	% (95% CI)	% (95% CI)
Diabetes mellitus	5.46 (4.78 to 6.15)	5.69 (5.05 to 6.33)	-2.25 (-1.13 to 6.84)	-3.96 (-19.65 to 11.73)
CVDs*	19.15 (17.88 to 20.42)	25.09 (23.68 to 26.51)	-5.94 (-7.68 to -4.20)	-23.68 (-29.79 to -17.57)
COPD*	4.01 (3.45 to 4.56)	2.82 (2.35 to 3.29)	1.19 (0.47 to 1.90)	42.10 (11.85 to 72.35)
Asthma*	4.36 (3.77 to 4.96)	1.64 (1.29 to 1.99)	2.72 (2.05 to 3.39)	165.82 (99.15 to 232.49)
Parkinson's disease	0.50 (0.28 to 0.71)	0.38 (0.21 to 0.55)	0.11 (-0.16 to 0.39)	29.77 (-50.95 to 110.49)
Epilepsy	0.94 (0.64 to1.24)	1.33 (1.03 to 1.68)	-0.38 (-0.80 to 0.03)	-28.98 (-55.85 to 2.13)
Thyroid disorders	5.89 (5.20 to 6.58)	5.43 (4.78 to 6.08)	0.46 (-0.49 to 1.42)	8.57 (-97.72 to 26.91)

**Denotes significant difference between BHIS prevalence en BCHI prevalence.*
CVDs = cardiovascular diseases (including hypertension)
COPD = chronic obstructive pulmonary disease
[a]*Computed before rounded the estimated prevalences*

The agreement measures are described in Table 4.2.3. Sensitivity was good for CVDs (83%), fair for diabetes (78%) and poor for all other CDs (value varying between 27% and 67%). The specificity was excellent for all CDs (specificity above 98%) except for CVDs (specificity = 89%). The PPV was poor to fair for all the CDs (PPV varying between 41% and 75%), except for Parkinson's disease where it was good (PPV = 83%). Irrespective of the CDs, the NPV was excellent (NPV varying between 96% and 99%). The Kappa statistic was good for diabetes, CVDs, Parkinson's disease and thyroid disorders (kappa between 0.63 and 0.77), moderate for epilepsy (kappa = 0.46) and fair for COPD and asthma (kappa = 0.35).

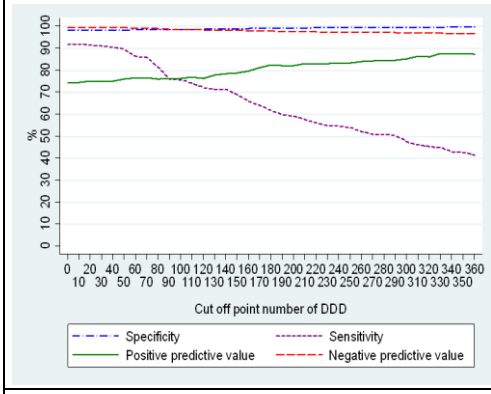
Table 4.2.3: Agreement between self-reported chronic disease and definitions from administrative data, HISlink 2013, Belgium*

Chronic disease	Sensitivity (%) (95% IC)	Specificity (%) (95% IC)	PPV (%) (95% IC)	NPV (%) (95% IC)	Kappa (95% CI)
Diabetes mellitus	78.5 (72.1-85.0)	98.5 (98.2-98.9)	75.4 (70.5-80.3)	98.8 (98.3-99.2)	0.77 (0.75-0.80)
CVDs	83.1 (80.6-85.6)	88.6 (87.6-89.7)	63.4 (60.6-66.2)	95.7 (95.0-96.3)	0.63 (0.61-0.65)
COPD	28.8 (22.3-35.3)	98.3 (97.9-98.6)	40.9 (32.5-49.3)	97.1 (96.6-97.5)	0.35 (0.30-0.40)
Asthma	27.4 (21.3-33.6)	99.5 (99.3-99.7)	72.9 (63.8-82.1)	96.8 (96.2-97.3)	0.35 (0.30-0.41)
Parkinson's disease	64.3 (38.9-89.8)	99.9 (99.9-100)	83.5 (69.8-97.3)	99.8 (99.7-100)	0.70 (0.58-0.82)
Epilepsy	60.4 (44.6-76.2)	99.2 (99.0-99.4)	42.9 (31.0-54.8)	99.6 (99.4-99.8)	0.46 (0.37-0.55)
Thyroid disorders	66.7 (61.0-72.4)	98.4 (98.1-98.7)	72.4 (67.3-77.5)	97.9 (97.5-98.3)	0.66 (0.62-0.69)

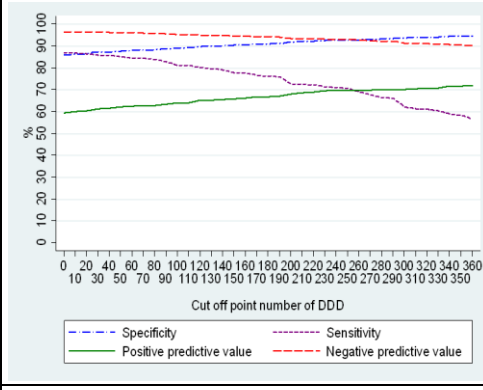
CVDs = cardiovascular diseases (including hypertension)
COPD = chronic obstructive pulmonary disease
PPV = positive predictive value; NPV = negative predictive value
** Sensitivity, specificity, PPV and NPV presented with self-reported as the referent*

The results of the sensitivity analysis are presented in Figure 4.2.1. Across the CDs, the sensitivity decreased with the increase of the cut-off point of the DDD, while the PPV slightly increased after the threshold of 90 DDD. Notable for Parkinson's disease was the highest PPV around 320 DDDs and for thyroid disorders was the lowest PPV around 220 DDDs.

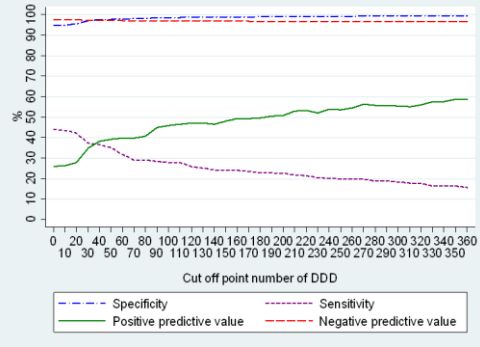
Diabetes



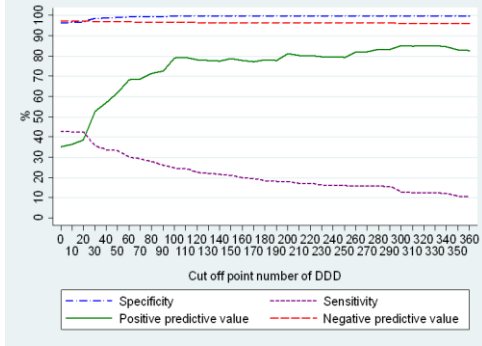
Cardiovascular diseases



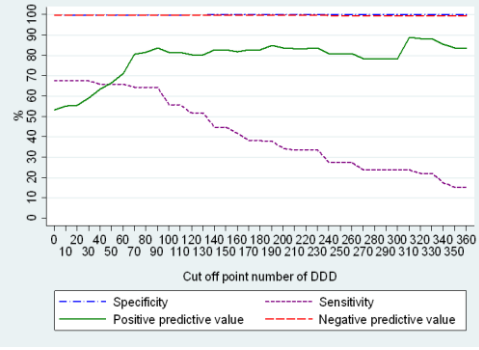
COPD



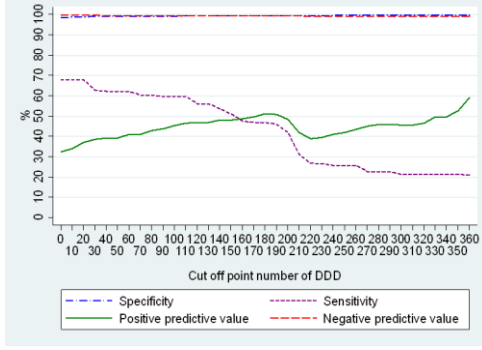
Asthma



Parkinson's disease



Epilepsy



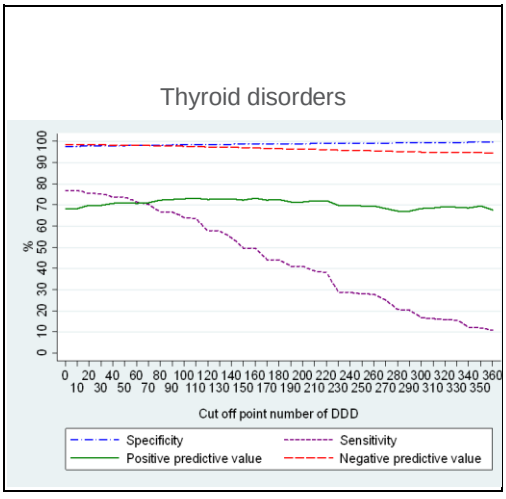


Figure 4.2.1: Sensitivity analysis: validity measures of chronic diseases as a function of the DDD threshold, HISlink 2013, Belgium

Table 4.2.4 show the results from the multivariable logistic regression, while the unadjusted odds ratios are presented in additional Table A2 (supplementary material). Table 4.2.4 shows that the agreement between BHIS and BCHI data sources is affected by individual sociodemographic characteristics and health status. However, the characteristics which are associated, the magnitude and direction of the effect varied across CDs. For instance, gender was not significantly associated with the agreement between BHIS and BCHI data except for thyroid disorders where the agreement was significantly lower among women (OR: 0.26, 95% CI: 0.17-0.40). Compared to the reference age-group (55-74 years), belonging to the youngest age-group (15-34 years) was associated with a greater level of agreement between the data sources for diabetes (OR: 6.40, 95% CI: 2.38-17.25), CVDs (OR: 8.63, 95% CI: 5.56-13.39) and thyroid disorders (OR: 2.76, 95% CI: 1.54-4.95), while the reverse is true for asthma (OR: 0.19, 95% CI: 0.10-0.36). Regarding participant's health status, people with a relatively good subjective health, those without multimorbidity and those who didn't simultaneous use 5 medicines or more on a typical day (polypharmacy) have greater odds of agreement between the two sources except for CVDs where the absence of multimorbidity was significantly associated with a lower odds of agreement.

Table 4.2.4: Odds Ratios[a] (95% CIs) for predictors of agreement between administrative and survey data for chronic diseases, HISlink 2013, Belgium

	Diabetes	CVDs	COPD	Asthma	Thyroid disorders
Gender					
Male	Ref.	Ref.	Ref.	Ref.	Ref.
Female	1.21 (0.71-2.06)	0.92 (0.73-1.15)	0.89 (0.62-1.27)	1.28 (0.88-1.86)	0.26 (0.17-0.40)*
Age group					
15-34	6.40 (2.38-17.25)*	8.63 (5.56-13.39)*	1.19 (0.55-2.56)	0.19 (0.10-0.36)*	2.76 (1.54-4.95)*
35-54	1.09 (0.56-2.10)	2.02 (1.50-2.72)*	0.81 (0.51-1.26)	0.51 (0.30-0.87)*	1.59 (0.98-2.58)
55-74	Ref.	Ref.	Ref.	Ref.	Ref.
75+	1.16 (0.54-2.47)	0.47 (0.35-0.63)*	0.98 (0.61-1.55)	2.09 (1.13-3.84)*	1.01 (0.60-1.69)
Education					
Low	Ref.	Ref.	Ref.	Ref.	Ref.
Intermediate	0.81 (0.42-1.57)	1.19 (0.89-1.60)	1.13 (0.75-1.70)	1.72 (1.08-2.74)*	1.08 (0.70-1.72)
High	1.55 (0.78-3.10)	1.08 (0.76-1.51)	2.07 (1.30-3.32)*	1.52 (0.89-2.59)	1.24 (0.75-2.06)
Nationality					
Belgian	2.57 (0.60-10.98)	0.76 (0.26-2.24)	0.28 (0.09-0.83)*	0.64 (0.29-1.42)	2.43 (0.95-6.25)
EU-countries	2.82 (0.52-15.35)	1.37 (0.43-4.32)	0.36 (0.10-1.29)	0.61 (0.18-2.05)	7.15 (1.68-30.51)*
Other countries	Ref.	Ref.	Ref.	Ref.	Ref.
Income					
Quintile 1	Ref.	Ref.	Ref.	Ref.	Ref.
Quintile 2	0.63 (0.31-1.27)	0.99 (0.70-1.39)	0.92 (0.55-1.55)	0.94 (0.54-1.63)	0.61 (0.36-1.06)
Quintile 3	1.29 (0.53-3.17)	0.99 (0.69-1.41)	1.17 (0.68-2.02)	1.05 (0.60-1.84)	0.6 (0.36-1.14)
Quintile 4	1.33 (0.64-2.77)	1.25 (0.86-1.81)	1.11 (0.61-2.03)	1.29 (0.69-2.40)	0.69 (0.37-1.28)
Quintile 5	1.05 (0.43-2.51)	1.16 (0.77-1.74)	1.34 (0.69-2.57)	0.70 (0.35-1.40)	1.37 (0.69-2.72)
Region					
Flanders	1.25 (0.74-2.09)	1.27 (1.02-1.59)*	1.64 (1.13-2.39)*	1.82 (1.20-2.75)*	2.50 (1.72-3.64)*
Brussels	1.70 (0.84-3.44)	1.06 (0.79-1.43)	1.29 (0.83-1.99)	1.03 (0.62-1.72)	2.30 (1.41-3.75)*
Wallonia	Ref.	Ref.	Ref.	Ref.	Ref.
Perceived health					
Good to very good	0.97 (0.57-1.65)	1.64 (1.26-2.14)*	1.76 (1.23-2.54)*	1.61 (1.10-2.36)*	1.10 (0.71-1.70)
Very bad to fair	Ref.	Ref.	Ref.	Ref.	Ref.
Multimorbidity					
Yes	Ref.	Ref.	Ref.	Ref.	Ref.
No	5.97 (3.06-11.67)*	0.47 (0.32-0.71)*	6.22 (3.86-10.03)*	15.40 (9.40-25.22)*	1.02 (0.61-1.71)
Polypharmacy					
Yes	Ref.	Ref.	Ref.	Ref.	Ref.
No	1.43 (0.77-2.68)	2.03 (1.42-2.90)*	1.26 (0.81-1.98)	0.70 (0.43-1.14)	2.59 (1.46-4.60)*

*CVDs = cardiovascular diseases (including hypertension) ; COPD = chronic obstructive pulmonary disease; * Denotes significant difference between this group and the reference group. [a] adjusted for all other variables*

4.2.5. Discussion

In this study we assessed agreement between population-based administrative and survey data for ascertaining cases of diabetes, asthma, chronic obstructive pulmonary

disease, cardiovascular diseases (including hypertension), Parkinson's disease, thyroid disorders and epilepsy, for which BHIS data served as the gold standard. We also investigated the individual characteristics that could influence the agreement between both data sources.

Using the two data sources, we obtained inconsistent prevalence estimates in 3 out of the 7 CDs studied. Specifically, in CVDs (including hypertension), the prevalence was significantly higher in the BCHI data than in the BHIS data, while the inverse was true for COPD and asthma. The high prevalence of CVDs (including hypertension) according to the BCHI source (25%) compared to the BHIS prevalence (19%) could be explained by the use of drugs in this ATC group for other problems such as a high serum cholesterol for example. Some drugs may be assigned to two chronic diseases simultaneously, for example, beta-blockers are prescribed both for patients with hypertension and in patients with heart problems. As mentioned by Huber et al. in their study, an unique assignment of ATC-codes to heart diseases is challenging, and with the new trends in the use of various drugs for cardiac and hypertensive patients, a clear distinction between ATC-codes for cardiac diseases and hypertension is infeasible [9]. Therefore, we included hypertension in the BHIS based case definition of CVDs. The low prevalence of COPD and asthma in the administrative data could be explained by the fact that some people suffering from asthma or COPD do not necessarily take medications or less than 90 DDDs per year.

The estimated prevalence rate of diabetes mellitus from BCHI data is comparable to the one estimated in similar studies using health administrative database [9,10,22,23], but higher than those in others comparable studies [5,13]. Moreover, the prevalence of the respiratory illness (COPD, asthma) from BCHI is also comparable to those in similar in Netherlands, Italy and Swedish [5,13,24,25]. Regarding the prevalence of Parkinson disease, thyroid disorders and Epilepsy, our results are in line with those reported by Francesco Chini et al. in Italy using a prescribed database [13] and by Huber et al. in Switzerland using medical and pharmacy claims data [9]. Considering the CVDs (including hypertension), our estimated prevalence was lower than the prevalence obtained by Huber et al. (29%) based on pharmacy data [9]. This difference could be explained by the CDs case definition used in their study: people were considered as having CD if they have at least one prescription in one of the generated ATC-groups CDs at the end of the reference year, while our definition was

more selective (at least 90 DDDs per year which could correspond to several prescriptions (if small package) or more or less 3 months treatment per the given year.

We found that sensitivity of administrative CDs was good-to-fair for diabetes and CVDs and poor for the remaining CDs. Not surprisingly, the lowest sensitivity was for COPD and asthma. The sensitivity drop with the increase of the cut-off point of DDD, while the PPV increase.

CDs that are more prevalent or that are symptom-based may also be more reliably self-reported [26]. In our definition of CVDs in BHIS data source, we included hypertension, which may have contributed to increase the agreement between both data sources for CVDs.

The lower sensitivity of asthma (27.4%) in contrast with its relatively higher PPV (72.9%) in this study could be explained by the fact that most of the people suffering from a less severe case of asthma could not take up to 90 DDDs of the specific medication per year and those who reach that cut-off are certainly positive cases. Furthermore, in an exploratory analysis (results not shown), we found that 3 persons out of 10 suffering from this CD did not contact a health care professional in the past 12 months for that condition.

The agreement between the two data sources varies by participants' sociodemographic characteristics and health status. However, this moderating effect varies in magnitude across CDs. Our results are consistent with findings in previous studies [3,8,27]. For instance, Lix et al. found that agreement between self-reported and medical records of chronic conditions was higher among younger age-groups and in the absence of comorbidity [3].

This study presents a number of strengths that deserve to be highlighted. First, the large sample size and the use of comprehensive administrative data, covering 99% of the Belgian population. It should be noted that not all countries have the opportunity to have such data. Thus, the existence of rich and detailed health insurance administrative data covering almost the entire population constitutes an added value for public health research in Belgium. Second, we calculated five agreement measures to enable comparison between data sources. Third, using individual record linkage, we further examined predictors that could affect the agreement between both data sources.

A number of limitations should also be acknowledged. One of the main limitations is that the case definition of CDs in the administrative data source was based on prescription drug codes dispensed in public pharmacies only and therefore drugs dispensed in the hospital settings were not included. Another limitation is the lack of additional information such as ICD-10 codes or other clinical diagnostic codes in the case ascertainment from administrative data source. Indeed, validation studies often include information from various sources in the algorithms: health surveys, ICD-10 codes, ATC codes, other clinical diagnostic codes, etc., and this provides much better measures of agreement [2,3,7,10]. Finally, the BHIS data was used as the gold standard in this study because next to administrative data, it is the only source for obtaining population-based chronic disease prevalence estimates in Belgium. We acknowledged that self-reported data may not be an unbiased gold standard due to the risk of under-reporting or over-reporting of some chronic diseases. However, self-reported data have been used in previous studies to assess the validity of health administrative databases [20,28,29] and have shown higher agreement between these sources for chronic diseases that are more familiar to patients, well defined and require ongoing management [3,20,28,30,31]. Keeping this in mind, the CDs discussed in this study are sufficiently well known and defined that the risk of providing erroneous information from BHIS participants is negligible. Moreover, several studies have assessed the specificity of self-reported CDs compared to clinical diagnoses or medical records and have found that the specificity was at least 80% for asthma, hypertension, severe heart disease or heart attack, stroke, diabetes mellitus, epilepsy, and Parkinson's disease [32].

4.2.6. Conclusions

In conclusion, NHIDI’s algorithms are an acceptable alternative for the identification of cases of diabetes, cardiovascular diseases (without distinction of the different pathologies), Parkinson's disease and thyroid disorders. On the basis of the current definition of CDs from BCHI data source, there is a significant underestimation of the number of patients suffering from asthma and COPD. Further research is needed to refine the definitions of CDs from administrative data by using other comparators (clinical data, data from general practitioners such as the Intego registry) or using different thresholds to enhance NIHDI algorithms.

4.2.7. Bibliography

1. Chronic disease & Policy - European chronic disease alliance [Internet]. [cited 2020 Jun 4]. https://alliancechronicdiseases.org/chronic-disease-policy/ . Accessed 27 August 2020.
2. Smidth M, Sokolowski I. Developing an algorithm to identify people with Chronic Obstructive Pulmonary Disease (COPD) using administrative data. 2012;7.
3. Lix L, Shaw S, Burchill C, Metge C, Bond R. Population-based data sources for chronic disease surveillance. Chronic Diseases in Canada. 2008;29:8.
4. CONSTANCES-Diab Group, Fuentes S, Cosson E, Mandereau-Bruno L, Fagot-Campagna A, Bernillon P, et al. Identifying diabetes cases in health administrative databases: a validation study based on a large French cohort. Int J Public Health. 2019;64:441–50.
5. Slobbe LCJ, Füssenich K, Wong A, Boshuizen HC, Nielen MMJ, Polder JJ, et al. Estimating disease prevalence from drug utilization data using the Random Forest algorithm. European Journal of Public Health. 2019;29:615–21.
6. Vaes B, Ruelens C, Saikali S, Smets A, Henrard S, Renard F, et al. Estimating the prevalence of diabetes mellitus and thyroid disorders using medication data in Flanders, Belgium. European Journal of Public Health. 2018;28:193–8.
7. Gothe H, Rajsic S, Vukicevic D, Schoenfelder T, Jahn B, Geiger-Gritsch S, et al. Algorithms to identify COPD in health systems with and without access to ICD coding: a systematic review. BMC Health Serv Res. 2019;19:737.
8. Koller KR, Wilson AS, Asay ED, Metzger JS, Neal DE. Agreement Between Self-Report and Medical Record Prevalence of 16 Chronic Conditions in the Alaska EARTH Study. J Prim Care Community Health. 2014;5:160–5.
9. Huber CA, Szucs TD, Rapold R, Reich O. Identifying patients with chronic conditions using pharmacy data in Switzerland: an updated mapping approach to the classification of medications. BMC Public Health. 2013;13:1030.
10. Orueta JF, Nuño-Solinis R, Mateos M, Vergara I, Grandes G, Esnaola S. Monitoring the prevalence of chronic conditions: which data should we use? BMC Health Serv Res. 2012;12:365.
11. Walraven C van. A comparison of methods to correct for misclassification bias from administrative database diagnostic codes. International Journal of Epidemiology. 2018;47:605–16.
12. Fosse-Edorh S, Rigou A, Morin S, Fezeu L, Mandereau-Bruno L, Fagot-Campagna A. Algorithmes basés sur les données médico-administratives dans le champ des maladies endocriniennes, nutritionnelles et métaboliques, et en particulier du diabète. Revue d'Épidémiologie et de Santé Publique. 2017;65:S168–73.

13. Chini F, Pezzotti P, Orzella L, Borgia P, Guasticchi G. Can we use the pharmacy data to estimate the prevalence of chronic conditions? a comparison of multiple data sources. BMC Public Health. 2011;11:688.
14. IMA Atlas [Internet]. http://atlas.aim-ima.be/base-de-donnees. Accessed 27 August 2020.
15. EPS R13 - FLAGS Release 20190201 FR.pdf. https://aim-ima.be/IMG/pdf/eps_r13_-_flags_release_20190201_fr_-_vs2.pdf. Accessed 27 August 2020.
16. Demarest S, Van der Heyden J, Charafeddine R, Drieskens S, Gisle L, Tafforeau J. Methodological basics and evolution of the Belgian health interview survey 1997–2008. Arch Public Health. 2013;71:24.
17. AIM-IMA [Internet]. https://aim-ima.be/Donnees-141. Accessed 27 August 2020.
18. World health Organization (last). WHO Collaborating Centre for Drug Statistics Methodology: ATC classification index with DDDs and Guidelines for ATC classification and DDD assignment [Internet]. Oslo, Norway:; 2006. https://www.whocc.no/atc_ddd_index_and_guidelines/guidelines/. Accessed 27 August 2020.
19. Gary W. Oehlert. A Note on the Delta Method: The American Statistician: Vol 46, No 1. The American Statistician. 46:6.
20. Lix L, Yogendran M, Mann J. Defining and validating chronic diseases: an administrative data approach An Update with ICD-10-CA [Internet]. 2008 Nov. Available from: http://umanitoba.ca/faculties/health_sciences/medicine/units/chs/departmental_units/mchp/projects/media/ICD10_Final.pdf
21. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33:159.
22. Stock SAK, Redaelli M, Wendland G, Civello D, Lauterbach KW. Diabetes-prevalence and cost of illness in Germany: a study evaluating data from the statutory health insurance in Germany. Diabet Med. 2006;23:299–305.
23. de Lagasnerie G, Aguadé A-S, Denis P, Fagot-Campagna A, Gastaldi-Menager C. The economic burden of diabetes to French national health insurance: a new cost-of-illness method based on a combined medicalized and incremental approach. Eur J Health Econ. 2018;19:189–201.
24. Weidinger P, Nilsson JLG, Lindblad U. Medication prescribing for asthma and COPD: a register-based cross-sectional study in Swedish primary care. BMC Fam Pract. 2014;15:54.
25. on behalf of the “CRD Real-World Evidence” scientific board, Biffi A, Comoretto R, Arfè A, Scotti L, Merlino L, et al. Can healthcare utilization data reliably capture cases of chronic respiratory diseases? a cross-sectional investigation in Italy. BMC Pulm Med. 2017;17:20.

26. Corser W, Sikorskii A, Olomu A, Stommel M, Proden C, Holmes-Rovner M. Concordance between comorbidity data from patient self-report interviews and medical record documentation. BMC Health Serv Res. 2008;8:85.
27. Martin LM, Leff M, Calonge N, Garrett C, Nelson DE. Validation of Self-Reported Chronic Conditions and Health Services in a Managed Care Population. :4.
28. Muggah E, Graves E, Bennett C, Manuel DG. Ascertainment of chronic diseases using population health data: a comparison of health administrative data and patient self-report. BMC Public Health. 2013;13:16.
29. Singh JA. Accuracy of Veterans Affairs Databases for Diagnoses of Chronic Diseases. 2009;6:11.
30. Okura Y, Urban LH, Mahoney DW, Jacobsen SJ, Rodeheffer RJ. Agreement between self-report questionnaires and medical record data was substantial for diabetes, hypertension, myocardial infarction and stroke but not for heart failure. Journal of Clinical Epidemiology. 2004;57:1096–103.
31. Nooney JG, Kirkman MS, Bullard KM, White Z, Meadows K, Campione JR, et al. Identifying optimal survey-based algorithms to distinguish diabetes type among adults with diabetes. Journal of Clinical & Translational Endocrinology. 2020;21:100231.
32. Van der Heyden J, De Bacquer D, Tafforeau J, Van Herck K. Reliability and validity of a global question on self-reported chronic morbidity. J Public Health. 2014;22:371–80.

Berete et al. Archives of Public Health (2020) 78:136
https://doi.org/10.1186/s13690-020-00521-z

Archives of Public Health

CORRECTION Open Access

Correction to: Comparing health insurance data and health interview survey data for ascertaining chronic disease prevalence in Belgium

Finaba Berete[1,2*], Stefaan Demarest[1], Rana Charafeddine[1], Olivier Bruyère[3] and Johan Van der Heyden[1]

Correction to: Arch Public Health 78, 120 (2020)
https://doi.org/10.1186/s13690-020-00500-4

The original publication of this article [1] contained an error in the footnotes of Table 1 and an error in *the absolute difference for Diabetes mellitus* in Table 2 which were introduced during the publication process. The full tables are available via the original publication. In this correction article the incorrect and correct footnotes are shown. These errors do not alter the results and conclusion of the article. The original article has been updated.

Table 1:

Incorrect:

– [a] For people aged < = 50 years
– [b] For people aged > 50 years

Correct:

– [a] For people aged > 50 years
– [b] For people aged < = 50 years

Table 2:

Incorrect

Chronic disease	Absolute difference[b](E1-E2) % (95% CI)
(Continued)	
Diabetes mellitus	−2.25 (−1.13 to 6.84)

Correct

Chronic disease	Absolute difference[b](E1-E2) % (95% CI)
Diabetes mellitus	−0.23 (−1.13 to 0.68).

The original article can be found online at https://doi.org/10.1186/s13690-020-00500-4.

* Correspondence: Finaba.berete@sciensano.be
[1]SD Epidemiology and public health, Sciensano, Juliette Wytsmanstraat, 14 1050 Brussels, Belgium
[2]Department of Public Health, Epidemiology and Health Economics, University of Liège, Liège, Belgium
Full list of author information is available at the end of the article

Author details
[1]SD Epidemiology and public health, Sciensano, Juliette Wytsmanstraat, 14 1050 Brussels, Belgium. [2]Department of Public Health, Epidemiology and Health Economics, University of Liège, Liège, Belgium. [3]WHO Collaborating Centre for Public Health aspects of musculoskeletal health and ageing, Department of Public Health, Epidemiology and Health Economics, University of Liege, Liège, Belgium.

Published online: 21 December 2020

Reference

1. Berete F, Demarest S, Charafeddine R, et al. Comparing health insurance data and health interview survey data for ascertaining chronic disease prevalence in Belgium. Arch Public Health. 2020;78:120 https://doi.org/10.1186/s13690-020-00500-4.

4.3. ASSESSING POLYPHARMACY IN THE OLDER POPULATION: COMPARISON OF A SELF-REPORTED AND PRESCRIPTION BASED METHOD

The findings of this paper were published as:

Van der Heyden J, **Berete F**, Renard F, Vanoverloop J, Devleesschauwer B, De Ridder K and Bruyère O. Assessing polypharmacy in the older population: Comparison of a self-reported and prescription based method. *Pharmacoepidemiology and Drug Safety* 30.12 (2021): 1716-1726.

Received: 26 October 2020 | Revised: 2 June 2021 | Accepted: 29 June 2021
DOI: 10.1002/pds.5321

ORIGINAL ARTICLE

WILEY

Assessing polypharmacy in the older population: Comparison of a self-reported and prescription based method

Johan Van der Heyden[1] | Finaba Berete[1,2] | Françoise Renard[1] | Johan Vanoverloop[3] | Brecht Devleesschauwer[1,4] | Karin De Ridder[1] | Olivier Bruyère[5]

[1]Department of Epidemiology and Public Health, Sciensano, Brussels, Belgium
[2]Department of Public Health, Epidemiology and Health Economics, University of Liège, Liège, Belgium
[3]Intermutualistic Agency (IMA-AIM), Brussels, Belgium
[4]Department of Veterinary Public Health and Food Safety, Ghent University, Merelbeke, Belgium
[5]WHO Collaborating Centre for Public Health Aspects of Musculoskeletal Health and Ageing, Department of Public Health, Epidemiology and Health Economics, University of Liège, Liège, Belgium

Correspondence
Johan Van der Heyden, Sciensano, Juliette Wytsmanstraat 14, 1050 Brussels, Belgium.
Email: johan.vanderheyden@sciensano.be

Funding information
National Institute for Health and Disability Insurance; Federal and Inter-Federated Belgian Public Health authorities

Abstract

Purpose: To explore differences in the prevalence and determinants of polypharmacy in the older general population in Belgium between self-reported and prescription based estimates and assess the relative merits of each data source.

Methods: Data were used from participants aged ≥65 years of the Belgian national health survey 2013 (n = 1950). Detailed information was asked on the use of medicines in the past 24 h and linked with prescription data from the Belgian compulsory health insurance (BCHI). Agreement between polypharmacy (use or prescription ≥5 medicines) and excessive polypharmacy (≥10 medicines) between both sources was assessed with kappa statistics. Multinomial logistic regression was used to study determinants of moderate (5–9 medicines) and excessive polypharmacy (≥10 medicines) and over- and underestimation of prescription based compared to self-reported polypharmacy.

Results: Self-reported and prescription based polypharmacy prevalence estimates were respectively 27% and 32%. Overall agreement was moderate, but better in men (kappa 0.60) than in women (0.45). Determinants of moderate polypharmacy did not vary substantially by source of outcome indicator, but restrictions in activities of daily living (ADL), living in an institution and a history of a hospital admission was associated with self-reported based excessive polypharmacy only.

Conclusions: Surveys and prescription data measure polypharmacy from a different perspective, but overall conclusions in terms of prevalence and determinants of polypharmacy do not differ substantially by data source. Linking survey data with prescription data can combine the strengths of both data sources resulting in a better tool to explore polypharmacy at population level.

KEYWORDS
ageing, health survey, linkage, polypharmacy, population-based, prescription data

KEY POINTS

- The prevalence of self-reported and prescription based polypharmacy (simultaneous use or prescription ≥5 medicines) in the Belgian population ≥65 years is respectively 27% and 32%.
- There is a moderate agreement between the estimates from both sources, which is higher in men (kappa 0.60) than in women (kappa 0.45).

- Moderate polypharmacy is significantly associated with multimorbidity, an inpatient hospitalization in the past year and a higher number of contacts with the GP; excessive polypharmacy with lower secondary education, living in a nursing home, moderate and severe restrictions in activities of daily living and inpatient hospitalization in the past year.
- Health surveys in which detailed information is gathered on the use of medicines and prescription databases are complementary tools to study polypharmacy at population level.
- Linkages of survey data and prescription data offer new opportunities for research in the domain of polypharmacy.

4.3.1. Abstract

Purpose

To explore differences in the prevalence and determinants of polypharmacy in the older general population in Belgium between self-reported and prescription based estimates and assess the relative merits of each data source.

Methods

Data were used from participants aged ≥ 65 years of the Belgian national health survey 2013 (n = 1950). Detailed information was asked on the use of medicines in the past 24 hours and linked with prescription data from the Belgian compulsory health insurance. Agreement between polypharmacy (use or prescription ≥ 5 medicines) and excessive polypharmacy (≥ 10 medicines) between both sources was assessed with kappa statistics. Multinomial logistic regression was used to study determinants of moderate (5-9 medicines) and excessive polypharmacy (≥ 10 medicines) and over- and underestimation of prescription based compared to self-reported polypharmacy.

Results

Self-reported and prescription based polypharmacy prevalence estimates were respectively 27% and 32%. Overall agreement was moderate, but better in men (kappa 0.60) than in women (0.45). Determinants of moderate polypharmacy did not vary substantially by source of outcome indicator, but restrictions in activities of daily living, living in an institution and a history of a hospital admission was associated with self-reported based excessive polypharmacy only.

Conclusions

Surveys and prescription data measure polypharmacy from a different perspective, but overall conclusions in terms of prevalence and determinants of polypharmacy do not differ substantially by data source. Linking survey data with prescription data can combine the strengths of both data sources resulting in a better tool to explore polypharmacy at population level.

4.3.2. Background

The ageing of the population has led to an increase of multimorbidity in many countries[1–5]. From a systematic review of the literature it appears that the prevalence

of multimorbidity in older persons ranges from 55 to 98%[6]. For most chronic conditions there are disease-specific guidelines, including recommendations for the use of medicines to treat the disease or prevent complications. However, most clinical practice guidelines do not modify or discuss the applicability of their recommendations for older patients with multiple diseases and this inevitably leads to polypharmacy[7,8]. Polypharmacy can be appropriate, but is problematic when the increased risk of harm mainly due to drug-drug interactions and side effects outweighs plausible benefits[9].

Obtaining a clear and comprehensive picture of polypharmacy is a big challenge. Studies on polypharmacy vary with regard to the definition, but also by setting, reference period, age group of the study population, type, volume and regularity of use of medicines considered. Regarding definition, there are two approaches. A first one takes into account the quality of prescribing[10], but distinguishing appropriate and inappropriate polypharmacy remains difficult. A second approach advocates a definition based on the number of medications, but there is no theoretical basis that may confirm the number of medications required for such a definition[11]. A systematic review of numerical only definitions of polypharmacy found thresholds between ≥ 2 and ≥ 11, but the most commonly used approach is to define polypharmacy as the simultaneous use of 5 or more medicines on one day[10] and define excessive polypharmacy as the simultaneous use of 10 or more medicines.

Most population based studies on the use of medication are based on prescription data or self-reported survey data[12]. Prescription data might be more accurate as they are not prone to poor recall, but may not represent actual use. Often they are collected for reimbursement purposes and information on non-reimbursed medicines is lacking. Self-reported data (via a self-completed questionnaire, telephone interview, or face-to-face interview) provide information on the use of both prescribed and non-prescribed medicines. This can be supplemented by a medication inventory, whereby all medication packages are presented to interviewers, reducing any recall problems, as for instance is done in the National Health and Nutrition Examination Survey (NHANES), the Canadian Health Measures Survey (CHMS)[13], and the Belgian Health Interview Survey (BHIS)[14].

Comparison between prescription and self-reported data is essential for improved understanding of the relative merits of each source and the extent of potential misclassification of medication use in pharmacoepidemiological studies. It also adds

evidence on the reliability of epidemiologic studies that quantify medication use through self-report, which is often the easiest way to gather this type of information. The comparison of information on polypharmacy of prescription based and self-reported data is useful to understand strengths and weaknesses of both data sources and gain further insights on how to better interpret results from those data sources.

In this study, data linkage is used to compare simultaneous polypharmacy on a single day based on prescription data from the Belgian compulsory health insurance (BCHI) with a similar indicator based on the number of prescribed and non-prescribed medicines used in the past 24 hours according to the BHIS. The specific objectives of the study are 1) to assess to which extent polypharmacy and excessive polypharmacy are under- or overestimated if based on prescription data compared to reported use of medicines; 2) to explore differences and similarities in the estimates on the use of specific groups of medicines between prescription based and self-reported information; and 3) to investigate to which extent determinants of polypharmacy and excessive pharmacy in the older general population differ depending on the data source that was used to assess this.

4.3.3. Methods

Data

The BHIS is household suvey organized every 4 to 5 years. Participants are selected through a stratified clustered multistage sampling design[15]. The target population consists of all Belgian residents, including older people who live in nursing homes. In the BHIS, information is collected on the health status, health behavior, health care consumption and sociodemographic characteristics of all participants. As part of the Computer-Assisted Personal Interview (CAPI) respondents are asked to show to the interviewer the medicines they have used in the past 24 hours. The interviewer records the brand name of the medicine and if available the national code which can be found on the package. For each medicine it is asked whether it was taken on doctor's prescription or not and what was the reason to take the medicine. In a later stage information on the WHO Anatomical Therapeutic Chemical Classification (ATC) code and the reimbursement status is added by merging the data with information from the National Institute of Health and Disability Insurance.

BCHI data included comprehensive information on all reimbursed medicines for the years 2012, 2013 and 2014, more specifically: anonymized patient ID, date of prescription, national code of the medicine (with a direct link to the brand name), ATC code, quantity per package (QPP) and number of daily defined doses (DDD) per prescription.

For this study data were used of the BHIS 2013 participants aged 65 years and over. The participation rate of this survey at household level was 57,1%. Previous research showed that in the BHIS the participation rate of people aged 65 years and over is similar as in the younger age groups[16]. For 1,950 respondents (96.4% of the BHIS participants within this age group) data could be linked with prescription data on reimbursed medicines from the BCHI.

Outcome indicators and potential determinants

Polypharmacy status for both methods was classified into three groups: non-polypharmacy (< 5 medicines), moderate polypharmacy (5-9 medicines), and excessive polypharmacy (≥ 10 medicines). This classification has been used in the literature before[17]. Some analyses were also conducted on a binary polypharmacy indicator (≥ 5 medicines). Self-reported polypharmacy was defined taking into account all medicines included in the official Belgian compendium of medicines[18]. This is a comprehensive list of medicines available in Belgian public pharmacies, including both prescription medicines (reimbursed or not) and over-the-counter medicines (OTC). Herbal medicines, homeopathic medicines and most of the food supplements are not included in line with other studies[14,19]. Simultaneous polypharmacy on the date of the interview based on the BCHI data was calculated by the method proposed by Fincke et al.[20]. This method makes use of the date of dispensing of the medicine, the quantity per package (QPP) and the daily defined dose (DDD) to estimate if a medicine is “active” on a particular day, which means that the prescription is recent enough to assume that the person has been using this medicine on that day. In our study this method was applied to assess if a medicine was ‘active’ on the day of the interview.

Prescription data did not take into account non-reimbursed prescription medicines and OTC, because such information is not available in the BCHI database.

BHIS based potential determinants of polypharmacy status that were considered were gender, age, educational attainment, living situation, region of residence, multimorbidity, restrictions in activities of daily living (ADL), inpatient and day patient hospitalization in the past year and number of contacts with the general practitioner and the specialist in the past two months. Multimorbidity was defined as having suffered in the past year from at least two of the following diseases: serious heart disease, hypertension, obstructive lung disease, cancer, arthrosis or arthritis and diabetes. A similar survey-based multimorbidity indicator has been used in a Canadian study[21]. The ADL indicator in this study was based on questions on getting in and out of a bed or chair, dressing and undressing, bathing or showering, feeding yourself and using toilets from the European Health Interview Survey[22].

Statistical analyses

Agreement between self-reported and prescription based polypharmacy was assessed after having excluded important groups of medicines (in terms of use) which are always or usually OTC and/or not reimbursed (ATC G04CA, N02BE, N05CF, N05BA, N05CD, A12AX, M05BA) from both data sources. Using the self-reported based estimates as reference we calculated the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), with corresponding 95% confidence intervals, of prescription based polypharmacy (≥ 5 medicines) and excessive polypharmacy (≥ 10 medicines). The agreement between the estimates from both sources was assessed with kappa statistics, including 95% confidence intervals.

To gain further insights comparisons were also made between prevalence estimates of self-reported use and recent prescription of specific types of medicines at the ATC 4th level (ATC4), which corresponds in the ATC classification system with the chemical subgroup. This was done for the 25 ATC4 group categories that were most frequently reported and prescribed. These represent more than 70% of the total daily number of consumed and prescribed medicines, both in men and women.

For all groups of medicines, except the ATC groups mentioned above, statistically significant differences were assessed with the delta method[23]. In addition the agreement of both estimates was assessed with kappa statistics, including 95% confidence intervals.

In a subsequent step potential determinants of moderate and excessive polypharmacy were explored via odds ratios (OR) of multinomial logistic regression models. This was first done separately for prescription and self-reported based estimates. Then multinomial models were fitted to investigate potential determinants of under- and overestimation of the prescription versus the self-reported based estimate.

Analyses were conducted with SAS 9.4. and Stata 16.0 taking into account the design settings of the BHIS, including the survey weights, household clusters, and strata.

4.3.4. Results

Table 4.3.1 provides information on the distribution of the study sample by socio-demographic and health characteristics, before and after the application of survey weights.

Table 4.3.1: Description of the sample

		N	Crude percentage (sample)	Weighted percentage (population)
Gender				
	Men	858	44.0	42.3
	Women	1092	56.0	57.7
Age				
	65-74 years	998	51.2	50.1
	75-84 years	714	36.6	37.5
	85+ years	238	12.2	12.4
Education				
	No diploma/primary	491	25.2	27.8
	Lower secondary	391	20.1	20.3
	Higher secondary	499	25.6	26.6
	Tertiary	543	27.9	25.3
	No info	26	1.3	
Living situation				
	Alone	644	33.0	32.9
	At home with others	1203	61.7	63.1
	Institution	84	4.3	4.0
	Missing	19	1.0	
Region				
	Flanders	731	37.5	61.3
	Brussels	400	20.5	7.7
	Wallonia	819	42.0	31.0
Multimorbidity				
	Yes	683	35.0	35.9
	No	1263	64.8	64.1
	No info	4	0.2	
Restrictions in ADL*				
	Severe	291	14.9	16.0
	Moderate	256	13.1	12.2
	None	1402	71.9	71.7
	No info	1	0.1	

**Activities of daily living.*

The prescription based prevalence estimates of polypharmacy (≥ 5 medicines) and excessive polypharmacy (≥ 10 medicines) in the Belgian population aged 65 years and over are respectively 32,4% and 2.4%. Similar survey-based estimates based on the use of medicines are respectively 27,4% and 3.7% (Table 4.3.2). The match between self-reported and prescription based polypharmacy (≥ 5 medicines) is reasonable in men, with a sensitivity of 82.1%, a specificity of 84.6% and a kappa of 0.60. In women the agreement between self-reported and prescription based assessment of polypharmacy (≥ 5 medicines) is weaker (kappa 0.45). Table 4.3.2 further shows that there is a poor agreement between self-reported and prescription based excessive polypharmacy, which is again worse in women.

Table 4.3.2: Percentage of people aged 65 years and older with polypharmacy[1] and excessive polypharmacy[2] according to both sources and agreement[3] between both sources taking self-reported based data as reference

Polypharmacy[1]	Self-reported Based[4]		Prescription based[5]		Sensitivity		Specificity		PPV[6]		NPV[7]		Kappa[3]		n
	%	95% CI	%	95% CI	%	95% CI	%	95% CI	%	95% CI	%	95% CI		95%CI	
Men - 65-74 years	22.2	(17.5-26.9)	27.9	(22.7-33.1)	85.7	(77.2-94.3)	87.0	(82.7-91.4)	62.6	(52.0-73.2)	96.0	(93.5-98.5)	0.64	0.56-0.72	469
Men - 75-84 years	29.0	(23.0-35.1)	35.7	(28.7-42.7)	78.2	(68.2-88.2)	81.3	(73.6-88.9)	60.3	(47.3-73.4)	91.1	(86.7-95.6)	0.54	0.44-0.65	314
Men - 85 years +	34.3	(19.6-48.9)	44.2	(28.5-59.9)	82.7	(64.1-100.0)	80.9	64.1-97.8)	67.8	(41.1-94.6)	90.6	(80.7-100.0)	0.60	0.41-0.79	75
Men - all	25.7	(22.0-29.3)	32.1	(28.0-36.2)	82.4	(76.3-88.4)	84.6	(80.7-88.5)	62.3	(54.4-70.1)	93.9	(91.7-96.2)	0.60	0.54-0.66	858
Women - 65-74 years	24.0	(18.5-29.5)	27.8	(21.6-34.1)	70.8	(57.5-84.1)	85.1	(78.7-91.5)	54.5	(39.6-69.4)	92.0	(88.0-96.1)	0.50	0.42-0.59	529
Women - 75-84 years	31.7	(25.2-38.2)	34.8	(28.2-41.3)	66.2	(51.9-80.6)	80.1	(74.2-86.1)	55.4	(43.8-66.9)	86.5	(79.8-93.1)	0.44	0.35-0.53	400
Women - 85 years +	35.0	(25.9-44.0)	41.1	(31.4-50.8)	64.9	(47.9-81.9)	72.8	(61.5-84.2)	49.3	(32.2-66.4)	83.6	(74.8-92.4)	0.34	0.20-0.49	163
Women - all	28.6	(24.8-32.5)	32.6	(28.5-36.7)	67.7	(52.4-68.3)	81.5	(77.4-85.6)	53.9	(48.6-65.0)	88.8	(85.4-92.2)	0.45	0.40-0.47	1092
All	27.4	(24.6-30.2)	32.4	(29.4-35.3)	73.8	(68.3-79.4)	82.8	(79.8-85.8)	57.5	(51.5-63.5)	91.0	(88.7-93.2)	0.52	0.47-0.56	1950
Excessive polypharmacy[2]															
Men - 65-74 years	2.4	(0.8-4.1)	2.8	(1.2-4.5)	60.4	(4.3-100.0)	98.4	(97.1-99.6)	39.1	(5.9-72.2)	99.3	(98.2-100.0)	0.46	(0.19-0.74)	469
Men - 75-84 years	5.4	(2.6-8.3)	4.0	(1.4-6.6)	45.5	(10.8-80.3)	98.2	(96.3-100.0)	51.5	(10.2-92.8)	97.7	(95.8-99.6)	0.46	(0.20-0.72)	314
Men - 85 years +	9.0	(0.0-20.0)	1.2	(0.0-3.5)	32.4	(0.0-100.0)	100.0	-	100.0	-	97.5	(94.1-100.0)	0.48	(0.00-1.00)	75
Men – all	4.1	(2.4-5.8)	3.1	(1.8-4.4)	49.1	(24.3-73.9)	98.5	(97.5-99.4)	46.9	(23.1-70.7)	98.6	(97.6-99.5)	0.47	(0.28-0.65)	858
Women - 65-74 years	4.8	(1.1-8.5)	1.5	(0.2-2.8)	2.4	(0.0-8.5)	99.0	(98.1-99.9)	8.5	(0.0-28.9)	96.4	(92.8-100.0)	0.02	(0.00-0.13)	529
Women - 75-84 years	2.0	(0.6-3.3)	2.4	(0.9-3.9)	6.7	(0.0-25.2)	98.5	(97.4-99.6)	4.2	(0.0-14.3)	99.1	(98.2-100.0)	0.04	(0.00-0.22)	400
Women - 85 years +	3.4	(0.8-6.0)	1.2	(0.0-2.9)	36.0	(0.0-100.0)	100.0	-	100.0	-	98.8	(97.3-100.0)	0.53	(0.00-1.00)	163
Women – all	3.5	(1.7-5.3)	1.8	(1.0-2.7)	7.1	(0.0-18.0)	98.9	(98.3-99.5)	13.9	(0.0-32.5)	97.8	(96.1-99.5)	0.08	(0.00-0.21)	1092
All	3.7	(2.5-5.0)	2.4	(1.6-3.1)	26.3	(9.7-42.8)	98.7	(98.2-99.3)	34.8	(17.9-51.7)	98.1	(97.0-99.2)	0.28	(0.16-0.41)	1950

1 ≥ 5 medicines 2 ≥ 10 medicines

3 for the calculation of agreement measures important ATC groups which are always or usually OTC and/or not reimbursed were excluded from both data sources (G04CA, N02BE, N05CF, N05BA, N05CD, A12AX, M05BA)

4 not reimbursed prescription medicines and OTC (over-the-counter medicines) included; 5 not reimbursed prescription medicines and OTC not included

6 positive predictive value; 7 negative predictive value

Tables 4.3.3 and 4.3.4 present prevalence estimates of the use in the past 24 hours and a recent prescription for the 25 most frequently reported and prescribed ATC4 categories for which information is available in both databases.

Overall there is a good agreement between self-reported and prescription based estimates, with most kappas being higher than 0.50. For most ATC4 categories higher prevalence estimates are obtained for a recent prescription than for use in the past 24 hours

Table 4.3.3 Reported use of medicines in the past 24 hours versus recent prescription, by ATC4 code for the 25 most used and prescribed medicines, men aged 65 years and older

ATC4	Category	Reported use of a medicine in this category in past 24hrs (E1)		Recent prescription of a medicine in this category in BCHI° data (E2)		Absolute difference between both estimates (E2) - (E1)*		Agreement between both estimates	
		%	95%CI	%	95%CI	%	95%CI	kappa	95%CI
C10AA	Hydroxymethylglutaryl CoA reductase inhibitors (statins)	33.6	(29.3;37.8)	39.3	(34.9;43.7)	5.7	(-0.3;11.7)	0.63	(0.58-0.69)
B01AC	Platelet aggregation inhibitors excl. heparin	29.0	(25.2;32.9)	31.8	(27.7;35.8)	2.8	(-2.8;8.4)	0.61	(0.55-0.67)
C07AB	Beta blocking agents. selective	19.8	(16.4;23.3)	19.5	(15.9-23.1)	-0.3	(-5.3;4.6)	0.63	(0.56-0.69)
A02BC	Proton pump inhibitors	17.3	(13.7;20.8)	22.7	(18.6;26.7)	5.4*	(0.0;10.7)	0.69	(0.63-0.76)
C09AA	Angiotensin-converting enzyme inhibitors. plain	12.1	(9.3:15.0)	17.0	(13.8;20.3)	4.9*	(0.6;9.2))	0.72	(0.65-0.79)
C08CA	Dihydropyridine derivatives	10.5	(7.7:13.2)	14.9	(11.8;18.0)	4.4*	(0.3;8.6)	0.77	(0.71;0.84)
A10BA	Biguanides	8.9	(6.7:11.5)	8.8	(6.7;10.9)	-0.1	(-3.2;3.0)	0.63	(0.53-0.72)
G04CA	Alpha-adrenoreceptor antagonists	8.5	(6.0:10.9)	Not listed because only partial information in prescription database					
M04AA	Preparations inhibiting uric acid production	7.5	(5.4:9.6)	7.6	(5.3;9.9)	0.1	(-3.1:3.2)	0.64	(0.54;0.74)
N05BA	Benzodiazepine derivatives	7.0	(5.0:8.9)	Not listed because no information in prescription database					
C03CA	Sulfonamides, plain	5.6	(3.6:7.6)	6.5	(4.6:8.4)	0.9	(-1.9:3.7)	0.66	(0.55;0.77)
C09CA	Angiotensin II receptor blockers, plain	5.6	(3.6:7.6)	8.5	(6.1;11.0)	2.9	(-0.2:6.1)	0.74	(0.65;0.83)
C09DA	Angiotensin II receptor blockers and diuretics	4.9	(3.0:6.9)	5.8	(3.8;7.9)	0.9	(-1.9:3.7)	0.78	(0.68;0.88)
A10BB	Sulfonylureas	4.8	(3.1:6.5)	6.1	(4.0:8.1))	1.2	(-1.4:3.9)	0.78	(0.69;0.88)

B01AA	Vitamin K antagonists	4.2	(2.5:5.8)	4.5	(2.6;6.4)	0.3	(-2.2:2.9)	0.69	(0.56;0.81)
R03AK	Adrenergics in combination with corticosteroids or other drugs	4.0	(2.2:5.8)	4.7	(3.1;6.4)	0.8	(-1.7:3.2)	0.55	(0.41;0.69)
N06AX	Antidepressants	3.8	(2.0:5.5)	3.8	(2.0;5.6)	0.0	(-2.5:2.5)	0.52	(0.37-0.68)
C07BB	Beta blocking agents, selective, and thiazides	3.6	(1.8:5.3)	3.1	(1.4;4.8)	-0.4	(-2.9:2.0)	0.86	(0.76;0.96)
C01DX	Vasodilators used in cardiac diseases	3.5	(2.1:4.9)	5.7	(3.6;7.7)	2.2	(-0.3:4.6)	0.72	(0.60;0.83)
N02BE	Anilides	3.4	(1.9:4.9)	Not listed because only partial information in prescription database					
N06AB	Selective serotonin reuptake inhibitors	3.1	(1.8:4.3)	4.3	(2.8;5.8)	1.2	(-0.7:3.2)	0.79	(0.68;0.90)
N05CD	Benzodiazepine derivatives	3.1	(1.7:4.5)	Not listed because no information in prescription database					
C07AA	Beta blocking agents, non-selective	3.0	(1.8:4.3)	3.5	(2.0;5.1)	0.5	(-1.5:2.5)	0.78	(0.66;0.91)
H03AA	Thyroid hormones	2.3	(1.3:3.3)	2.5	(1.4;3.6)	0.2	(-1.3:1.7)	0.66	(0.49;0.83)
R03BB	Anticholinergics	2.2	(1.1:3.4)	2.2	(0.9;3.5)	0.0	(-1.7:1.7)	0.61	(0.42;0.80)
C09BA	Angiotensin-converting enzyme inhibitors and diuretics	2.0	(0.9:3.1)	3.7	(1.7:4.4)	1.7	(-0.2:3.6)	0.57	(0.40;0.74)
R05CB	Mucolytics	1.7	(0.9:3.1)	5.6	(3.7;7.6)	4.0*	(1.8:6.2)	0.32	(0.17;0.48)
H02AB	Glucocorticoids	1.3	(0.7:2.4)	3.6	(1.8;5.4)	2.3*	(0.4:4.3)	0.24	(0.06;0.42)
R03AC	Selective beta-2-adrenoreceptor agonists	0.9	(0.1:1,6)	3.2	(1.4;5.0)	2.3*	(0.3:4.3)	0.38	(0.18;0.59)
S01ED	Beta blocking agents (ophthalmological treatment)	0.7	(0.1:1.2)	2.7	(1.1;4.4)	2.0*	(0.3:3.8)	0.22	(0.02;0.43)

° Belgian Compulsory Health Insurance
** significant difference ($p < 0.05$)*

Table 4.3.4 Reported use of medicines in the past 24 hours versus recent prescription, by ATC4 code for the 25 most used and prescribed medicines°, women aged 65 years and older

ATC4	Category	Reported use of a medicine in this category in past 24hrs (E1)		Recent prescription of a medicine in this category in BCHI° data (E2)		Absolute difference between both estimates (E2) - (E1)*		Agreement between both estimates	
		% [a]	95%CI	% [b]	95%CI	%	95%CI	kappa	95%CI
C10AA	Hydroxymethylglutaryl CoA reductase inhibitors (statins)	30.8	(26.9;34.7)	39.3	(35.1;43.6)	8.5*	(2.7;14.3)	0.63	(0.58;0.68)
B01AC	Platelet aggregation inhibitors excl. heparin	20.4	(17.2:23.7)	25.7	(21.8;29.7)	5.3*	(0.2;10.4)	0.64	(0.59;0.70)
C07AB	Beta blocking agents. selective	17.9	(14.8:20.9)	17.3	(14.1;20.6)	-0.5	(-5.0;3.9)	0.56	(0.50;0.62)
A02BC	Proton pump inhibitors	15.5	(12.6:18.5)	24.4	(20.5;28.2)	8.8*	(4.0;13.6)	0.60	(0.54;0.66)
N05BA	Benzodiazepine derivatives	12.4	(10.0:14.8)	Not listed because no information in the prescription database					
H03AA	Thyroid hormones	11.5	(9.2:14.7)	11.9	(9.2;14.7)	0.4	(-3.2;4,1)	0.73	(0.67;0.80)
C09AA	Angiotensin-converting enzyme inhibitors. plain	9.1	(6.8:11.4)	14.8	(11.5-18.2)	5.7*	1.7;9.7)	0.65	(0.59;0.72)
A10BA	Biguanides	8.5	(5.9:11.0)	7.9	(5.3-10.4)	-0.6	(-4.2;3.0)	0.65	(0.57-0.73)
N06AB	Selective serotonin reuptake inhibitors	8.0	(5.5:10.5)	12.2	(9.4-15.1)	4.2	(0.5;8.0)	0.73	(0.66-0.79)
C08CA	Dihydropyridine derivatives	7.6	(5.7:9.6)	11.1	(8.7-13.5)	3.4	(0.3;6.5)	0.67	(0.60-0.75)
N05CD	Benzodiazepine derivatives	7.5	(5.1:9.8)	Not listed because no information in the prescription database					
N02BE	Anilides	7.3	(5.0:9.6)	Not listed because only partial information in the prescription database					
N06AX	Antidepressants	4.7	(3.1:6.2)	5.2	(3.5;6.8)	0.5	(-1.7;2.8)	0.71	(0.62-0.81)
C09DA	Angiotensin II receptor blockers and diuretics	5.4	(3.5:7.3)	5.8	(3.8-7.7)	0.3	(-2.4;3.1)	0.76	(0.68;0.85)
C09CA	Angiotensin II receptor blockers - plain	5.0	(3.0:7.0)	7.2	(5.0;9.4)	2.2	(-0.7;5.2)	0.74	(0.66-0.82)
C03EA	Low-ceiling diuretics and potassium-sparing agents	5.2	(3.3:7.1)	5.9	(4.0;7.7)	0.7	(-2.0;3;3)	0.66	(0.56-0.75)

C03CA	Sulfonamides. Plain	5.0	(2.6:7.5)	5.9	(4.1;7.7)	0.8	(-2.2;3.8)	0.55	(0.44;0.66)
C07BB	Beta blocking agents. selective. and thiazides	4.5	(2.6:6.4)	5.1	(3.2;7.0)	0.6	(-2.1;3.3)	0.73	(0.64;0.83)
A12AX	Calcium. combinations with vitamin D and/or other drugs	4.6	(2.8:6.4)	Not listed because only partial information in the prescription database					
C07AA	Beta blocking agents. non-selective	4.1	(2.6:5.7)	4.3	(2.8;5.8)	0.2	(-2.0;2.3)	0.74	(0.64;0.84)
R03AK	Adrenergics in combination with corticosteroids or other drugs	4.2	(1.8:6.5)	5.2	(2.8;7.7)	1.1	(-2.3;4.5)	0.82	(0.73-0.90)
N05CF	Benzodiazepine related drugs	4.0	(2.4:5.6)	Not listed because no information in the prescription database					
N02AX	Opioids	3.1	(2.1:4.2)	1.6	(0.8;2.3))	-1.6*	(-2.8; -0.3)	0.44	(0.28;0.61)
C01DX	Vasodilators used in cardiac diseases	3.4	(1.5:5.3)	4.9	(2.8;7.0)	1.5	(-1.3;4.4)	0.81	(0.72;0.90)
C09BA	Angiotensin-converting enzyme inhibitors and diuretics	3.5	(1.9:5.1)	4.0	(2.3;5.6)	0.4	(-1.8;2.7)	0.73	(0.62;0.84)
N07CA	Antivertigo preparations	3.0	(1.8:4.2)	4.1	(2.6;5.6)	1.1	(-0.8;3,1)	0.73	(0.62;0.84)
S01ED	Beta blocking agents	1.9	(0.6:3.2)	3.9	(2.4;5.3)	2.0*	(0.0;4;0)	0.39	(0.23;0.54)
R05CB	Mucolytics	1.8	0.9;5.6)	4.5	(3.0;6.0)	2.7*	(1.0:4.5)	0.42	(0.27;0.57)
M05BA	Bisphosphonates	1.3	(0.5:2.1)	4.8	(3.2;6.4)	3.5*	(1.7:5.3)	0.27	(0.13;0.41)
R06AE	Piperazine derivatives	1.3	(0.5:2.1)	3.9	(1.6;6.2)	2.6*	(0.1:5.0)	0.39	(0.23;0.55)
M05BB	Bisphosphonates. combinations	0.9	(0.3:1.4)	3.8	(2.3;5.4)	3.0*	(1.3:4.6)	0.31	(0.15;0.48)

° *Belgian Compulsory Health Insurance; * significant difference ($p < 0.05$)*

Results from the multinomial logistic regression analyses (Table 4.3.5) show that regardless of the source of the data used, moderate polypharmacy is significantly associated with multimorbidity, inpatient hospitalization in the past year and a higher number of contacts with the GP in the past two months.

There is a significant association between self-reported based excessive polypharmacy and lower secondary education, living in a nursing home, moderate and severe restrictions in ADL, and inpatient hospitalization in the past year. However, no such significant associations are found for prescription based excessive polypharmacy (Table 4.3.5).

Underestimation of polypharmacy status in older people (of prescription based compared to self-reported based estimates) occurs significantly more often in women, people with low education, multimorbidity, moderate restrictions in ADL and an inpatient hospitalization in the past year (Table 4.3.6). Overestimation of the polypharmacy status (of prescription based compared to self-reported based estimates) is significantly associated with multimorbidity and a higher number of contacts with the GP in the past 2 months.

Table 4.3.5: Determinants of moderate (5-9 medicines) and excessive polypharmacy (≥ 10 medicines) in the population aged 65 years and older. Results from multinomial logit models

		Self-reported based estimates[1]		Prescription based estimates[2]	
		Moderate polypharmacy	Excessive polypharmacy	Moderate polypharmacy	Excessive polypharmacy
		OR[3] (+95% CI)	OR[3] (+95% CI)	OR[3] (+95% CI)	OR[3] (+95% CI)
Female		0.97 (0.72-1.31)	0.73 (0.32-1.66)	0.86 (0.63-1.16)	0.49 (0.21-1.17)
Age	65-74 years	Ref	Ref	Ref	Ref
	75-84 years	1.29 (0.91-1.82)	0.61 (0.28-1.32)	1.27 (0.91-1.76)	1.17 (0.54-2.55)
	85+ years	1.04 (0.61-1.77)	0.45 (0.12-1.68)	1.54 (0.95-2.50)	0.43 (0.14-1.35)
Education	No diploma/primary	0.96 (0.59-1.55)	0.97 (0.39-2.37)	1.09 (0.69-1.73)	0.74 (0.29-1.91)
	Lower secondary	1.51 (0.94-2.41)	3.11 (1.01-9.60)*	1.19 (0.77-1.84)	2.03 (0.78-5.30)
	Higher secondary	0.80 (0.51-1.26)	0.84 (0.32-2.21)	1.13 (0.75-1.73)	1.61 (0.59-4.41)
	Tertiary	Ref	Ref	Ref	Ref
Living situation	At home with others	Ref	Ref	Ref	Ref
	At home alone	1.30 (0.91-1.88)	0.53 (0.22-1.31)	1.12 (0.80-1.57)	1.06 (0.44-2.57)
	In a nursing home	1.49 (0.70-3.17)	3.94 (1.14-13.60)*	1.15 (0.51-2.59)	1.71 (0.39-7.43)
Region	Flanders	Ref	Ref	Ref	Ref
	Brussels	1.18 (0.80-1.74)	0.91 (0.39-2.11)	0.97 (0.67-1.41)	1.67 (0.64-4.33)
	Wallonia	1.12 (0.80-1.57)	1.14 (0.61-2.11)	0.95 (0.70-1.30)	2.01 (0.96-4.24)
Multimorbidity		3.58 (2.60-4.94)*	4.42 (2.09-9.35)*	3.93 (2.83-5.44)*	7.35 (3.25-16.65)*
Restrictions in ADL[4]	No restrictions	Ref	Ref	Ref	Ref
	Moderate restrictions	1.30 (0.84-2.04)	3.47 (1.31-9.18)*	1.08 (0.70-1.69)	1.65 (0.64-4.23)
	Severe restrictions	1.61 (0.98-2.65)	4.74 (1.59-14.09)*	1.01 (0.63-1.61)	2.37 (0.93-6.05)
Inpatient hospitalisation < 1 year		1.63 (1.12-2.37)*	3.47 (1.35-8.92)*	1.87 (1.27-2.75)*	1.78 (0.84-3.80)
Day patient hospitalisation < 1 year		1.06 (0.67-1.69)	0.52 (0.17-1.59)	0.78 (0.50-1.21)	0.34 (0.11-1.06)
Number contacts general practitioner < 2 months		1.22 (1.04-1.42)*	1.18 (1.01-1.38)*	1.27 (1.09-1.50)*	1.30 (1.08-1.57)*
Number contacts specialist < 2 months		1.00 (0.94-1.07)	1.04 (0.95-1.14)	1.04 (0.95-1.14)	1.07 (0.96-1.20)

*[1] not reimbursed prescription medicines and OTC (over-the-counter medicines) included;[2] not reimbursed prescription medicines and OTC not included ;[3] odds ratio[4] activities of daily living * significant difference (p < 0.05)*

Table 4.3.6 Determinants of under- and overestimation of prescription based polypharmacy status[1] in the population aged 65 years and older (reference = self-reported based estimate[2]). Results from multinomial logit models

		Underestimation[3]	Overestimation[4]
		OR[5] (+95% CI)	OR[5] (+95% CI)
Female		1.56 (1.01-2.41)*	1.19 (0.79-1.78)
Age	64-74 years	Ref	Ref
	75-84 years	0.98 (0.60-1.61)	1.18 (0.77-1.81)
	85+ years	0.74 (0.33-1.66)	1.39 (0.74-2.61)
Education	No diploma/primary	1.84 (1.03-3.29)*	1.94 (1.08-3.51)*
	Lower secondary	2.74 (1.41-5.35)*	1.36 (0.77-2.41)
	Higher secondary	1.32 (0.76-2.31)	2.21 (1.28-3.84)*
	Tertiary	Ref	Ref
Living situation	At home with others	Ref	Ref
	At home alone	0.83 (0.48-1.43)	0.88 (0.56-1.38)
	In a nursing home	2.00 (0.81-4.98)	1.07 (0.38-3.03)
Region	Flanders	Ref	Ref
	Brussels	1.44 (0.87-2.37)	1.23 (0.77-1.98)
	Wallonia	1.29 (0.83-2.01)	1.17 (0.78-1.98)
Multimorbidity		2.10 (1.32-3.35)*	2.39 (1.55-3.68)*
Restrictions in ADL[5]	No restrictions	Ref	Ref
	Moderate restrictions	2.33 (1.37-3.96)*	1.22 (0.71-2.09)
	Severe restrictions	1.65 (0.85-3.19)	0.69 (0,39-1,23)
Inpatient hospitalisation < 1 year		1.99 (1.15-3.43)*	1.50 (0.89-2.53)
Day patient hospitalisation < 1 year		0.88 (0.47-1.64)	0.70 (0.37-1.30)
Number contacts general practitioner < 2 months		1.06 (0.96-1.16)	1.09 (1.00-1.19)*
Number contacts specialist < 2 months		0.99 (0.88-1.11)	1.00 (0.93-1.07)

[1] not reimbursed prescription medicines and OTC (over-the-counter medicines) not included
[2] not reimbursed prescription medicines and OTC included
[3] non-polypharmacy according to the prescription based definition and moderate/excessive polypharmacy according to the self-reported based definition OR moderate polypharmacy according to the prescription based definition and excessive polypharmacy according to the self-reported based definition
[4] excessive polypharmacy according to the prescription based definition and non-polypharmacy/moderate polypharmacy according to the self-reported based definition OR moderate/excessive polypharmacy according to the prescription based definition and non-polypharmacy according to the self-reported based definition
[5] odds ratio
[6] activities of daily living
** significant difference (p < 0.05)*

4.3.5. Discussion

To our knowledge this is the first study that assessed polypharmacy within the same population based sample comparing self-reported and prescription based estimates. Cautiousness is needed to interpret the results because the first data source also includes not reimbursed prescription medicines and OTC, whereas the second one reimbursed medicines only, but even when only comparable medication groups were considered, we found that overall agreement was moderate. Determinants of moderate polypharmacy and excessive polypharmacy did not vary substantially by source of outcome indicator. Differences in the classification of the polypharmacy status between the two sources were associated with education, health status and health care use.

In many countries there is a systematic collection of prescription data, often linked to the reimbursement of medicines. Many population based polypharmacy studies use such prescription data [24–29] , which are considered to be reliable. However, the value of prescription data to assess polypharmacy in the population depends on several factors, the most important ones being the completeness of the target population included in the database and the validity and degree of the completeness of the registered data[30]. An important disadvantage of most pharmacoepidemiological databases is the lack of information on OTC and prescription medicines not subsidized by the National Health Insurance. Furthermore, studies using prescription data do not take into account that due to non-compliance and the intermittent use of prescribed medicines in case of symptoms only (e.g. painkillers), prescription data will not always correctly reflect the actual use that causes the hazardous effects of polypharmacy, such as adverse drug reactions and drug-drug interactions [31,32].

Surveys collect information on the actual use of medicines, and a number of studies have used this information to study polypharmacy[33–36]. Associations between polypharmacy and multimorbidity, functional limitations, educational attainment and visits to physicians, observed in our study, were also found in these studies. However, whereas most of these studies showed a higher likelihood of polypharmacy with increasing age and female gender, this was not observed in our study. In the BHIS interviewers did a visual inspection of the brand names of the medicines that were consumed and the reference period was short (24 hours). For this reason and also

because hazardous effects of polypharmacy are very much related to drug-drug interactions following the actual consumption of medicines, we used the self-reported data as reference for the comparison analyses. As both methods assess polypharmacy in a different way no perfect match was expected. The fact that the self-reported based estimate of polypharmacy is somewhat lower than the prescription based estimate may be related to an underreporting of medicines in the survey, but also to an overestimation of simultaneous polypharmacy in the prescription data, and this despite the fact that non reimbursed medicines and OTC medicines are not included in the latter database. Possible hypotheses for gender differences with respect to the agreement between self-reported and prescription based polypharmacy are gender differences in therapeutic compliance and in the use of medicines which were prescribed earlier, hence not identified in the insurance database as “active medicines”.

To gain further insights into differences between reported use and recent prescription of particular groups of medicines, we compared this for the most commonly used ATC4 groups. Although a comparison at ATC5 group level (the chemical substance) is more relevant if the emphasis lies on the number of pills a patient takes and the problems/confusion that can go together with it, and a comparison at ATC3 group level (the therapeutic subgroup) more relevant if the emphasis lies on the interactions between different types of medication that can cause dizziness, confusion, delirium,.., we opted for the ATC4 group level (the chemical subgroup), because it was found that the simultaneous use of medicines belonging to the same therapeutic subgroup occurs regularly, whereas this is not the case for medicines of the same chemical subgroup, and our sample was too small to provide sufficiently accurate estimates of the use of medicines at the ATC5 group level. Our results were quite satisfactory, with moderate to good levels of agreement for most groups, which is compatible with other studies in which this was assessed[12,37,38].

It is remarkable that the significant associations between the self-reported based indicator of excessive polypharmacy and restrictions in ADL, living in an institution and a history of a hospital admission were not significant when using the prescription based estimate as outcome indicator. Other studies have found associations between polypharmacy and these factors[33,39–42]. Furthermore, self-reported excessive polypharmacy estimates are higher than estimates based on self-reports, which is

logical because also OTC and not reimbursed prescription medicines are considered. These findings suggest that the assessment of excessive polypharmacy is more accurate when it is based on self-reports.383

Our results further indicate that differences in the classification of the polypharmacy status (no polypharmacy/moderate polypharmacy/excessive polypharmacy) between self-reported and prescription based estimates vary by population group. A lower education and the presence of multimorbidity are associated with more discrepancies in the polypharmacy status between both sources. A hypothesis is that the validity of self-reported information in these population groups is probably weaker as a result of more incomplete reporting of the medicines that have been used in the past 24 hours by these groups. This needs to be taken into consideration when interpreting studies on determinants of polypharmacy based on survey data.

According to the self-reported information people in nursing homes and with restrictions in ADL have higher risk of excessive polypharmacy, but this association is not seen in the prescription data. This could be related to the fact that these people may have received more assistance form caregivers when completing the survey.

Strengths and limitations of the study

Strengths of this study are that is population-based, it includes nursing home residents and information on self-reported and prescription based use of medicines was obtained from the same individuals. Limitations are mainly related to the validity of the information that is obtained. Even though the self-reported indicator is based on medicines that are actually shown to the interviewer, the list can be incomplete due to reluctance of the respondent to disclose the use of particular medicines. Furthermore medicines could have been missed if they are not taken daily, but every other day. With respect to the prescription based indicator, Fincke's method to identify an active medicine on the date of the interview is a good approximation, but as the frequency and regularity of use and the dose per day is unknown, misclassification is possible.

Finally, in this study the definition of polypharmacy is based on the simple counting of medicines, without taking into account the reason and the regularity of the use/prescription, the specific medicines that are combined and the existing co-morbidity. Further population based studies on polypharmacy should focus on

inappropriate polypharmacy and explore and compare data sources which are most suitable to investigate this.

4.3.6. Conclusions

Both health surveys and prescription databases are useful instruments to asses polypharmacy in the general older population. In Belgium there is a reasonable agreement between the outcomes generated by both sources. Whereas determinants of moderate polypharmacy do not vary substantially by source of outcome indicator, self-reported based estimates seem to identify better than prescription based estimates in which population groups excessive polypharmacy occurs more often. From our study it is clear that each data source alone has advantages and limitations. Linkage of survey data, administrative databases and clinical databases will create opportunities to study polypharmacy at population level making use of more appropriate and relevant outcome indicators.

Linking survey data with prescription data can combine the strengths of both data sources resulting in a better tool to explore polypharmacy at population level.

4.3.7. Bibliography

1. Uijen AA, van de Lisdonk EH. Multimorbidity in primary care: prevalence and trend over the last 20 years. Eur J Gen Pract. 2008;14 Suppl 1:28-32. doi:10.1080/13814780802436093
2. King DE, Xiang J, Pilkerton CS. Multimorbidity Trends in United States Adults, 1988-2014. J Am Board Fam Med. 2018;31(4):503-513. doi:10.3122/jabfm.2018.04.180008
3. Lebenbaum M, Zaric GS, Thind A, Sarma S. Trends in obesity and multimorbidity in Canada. Prev Med. 2018;116:173-179. doi:10.1016/j.ypmed.2018.08.025
4. Li J, Green M, Kearns B, et al. Patterns of multimorbidity and their association with health outcomes within Yorkshire, England: baseline results from the Yorkshire Health Study. BMC Public Health. 2016;16:649. doi:10.1186/s12889-016-3335-z
5. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet. 2012;380(9836):37-43. doi:10.1016/S0140-6736(12)60240-2
6. Marengoni A, Angleman S, Melis R, et al. Aging with multimorbidity: A systematic review of the literature. Ageing Research Reviews. 2011;10(4):430-439. doi:10.1016/j.arr.2011.03.003
7. Mortazavi SS, Shati M, Keshtkar A, Malakouti SK, Bazargan M, Assari S. Defining polypharmacy in the elderly: a systematic review protocol. BMJ Open. 2016;6(3):e010989. doi:10.1136/bmjopen-2015-010989
8. Sönnichsen A, Trampisch US, Rieckert A, et al. Polypharmacy in chronic diseases-Reduction of Inappropriate Medication and Adverse drug events in older populations by electronic Decision Support (PRIMA-eDS): study protocol for a randomized controlled trial. Trials. 2016;17:57. doi:10.1186/s13063-016-1177-8
9. Mair A, Wilson M, Dreischulte T. Addressing the Challenge of Polypharmacy. Annu Rev Pharmacol Toxicol. 2020;60:661-681. doi:10.1146/annurev-pharmtox-010919-023508
10. Masnoon N, Shakib S, Kalisch-Ellett L, Caughey GE. What is polypharmacy? A systematic review of definitions. BMC Geriatr. 2017;17(1):230. doi:10.1186/s12877-017-0621-2
11. Sirois C, Laroche M-L, Guénette L, Kröger E, Cooper D, Émond V. Polypharmacy in multimorbid older adults: protocol for a systematic review. Syst Rev. 2017;6(1):104. doi:10.1186/s13643-017-0492-9
12. Richardson K, Kenny RA, Peklar J, Bennett K. Agreement between patient interview data on prescription medication use and pharmacy records in those aged older than 50 years varied by therapeutic group and reporting of indicated

health conditions. J Clin Epidemiol. 2013;66(11):1308-1316. doi:10.1016/j.jclinepi.2013.02.016

13. Hales CM, Servais J, Martin CB, Kohen D. Prescription Drug Use Among Adults Aged 40-79 in the United States and Canada. NCHS Data Brief. 2019;(347):1-8.
14. Walckiers D, Van der Heyden J, Tafforeau J. Factors associated with excessive polypharmacy in older people. Arch Public Health. 2015;73:50. doi:10.1186/s13690-015-0095-7
15. Demarest S, Van der Heyden J, Charafeddine R, Drieskens S, Gisle L, Tafforeau J. Methodological basics and evolution of the Belgian health interview survey 1997-2008. Arch Public Health. 2013;71(1):24. doi:10.1186/0778-7367-71-24
16. Van der Heyden J, Demarest S, Van Herck K, De Bacquer D, Tafforeau J, Van Oyen H. Association between variables used in the field substitution and post-stratification adjustment in the Belgian health interview survey and non-response. Int J Public Health. 2014;59(1):197-206. doi:10.1007/s00038-013-0460-7
17. Onder G, Liperoti R, Fialova D, et al. Polypharmacy in nursing home in Europe: results from the SHELTER study. J Gerontol A Biol Sci Med Sci. 2012;67(6):698-704. doi:10.1093/gerona/glr233
18. Répertoire. CBIP. Accessed August 30, 2020. https://www.cbip.be/fr/chapters
19. O'Dwyer M, Peklar J, McCallion P, McCarron M, Henman MC. Factors associated with polypharmacy and excessive polypharmacy in older people with intellectual disability differ from the general population: a cross-sectional observational nationwide study. BMJ Open. 2016;6(4). doi:10.1136/bmjopen-2015-010505
20. Fincke BG, Snyder K, Cantillon C, et al. Three complementary definitions of polypharmacy: methods, application and comparison of findings in a large prescription database. Pharmacoepidemiol Drug Saf. 2005;14(2):121-128. doi:10.1002/pds.966
21. Fortin M, Hudon C, Haggerty J, van den Akker M, Almirall J. Prevalence estimates of multimorbidity: a comparative study of two sources. BMC Health Services Research. 2010;10(1):111. doi:10.1186/1472-6963-10-111
22. Eurostat. European Health Interview Survey (EHIS Wave 2). Methodological Manual. European Union; 2013.
23. Oehlert G. A Note on the Delta Method. Am Statitician. 1992;46:27-29.
24. Baek Y-H, Shin J-Y. Trends in polypharmacy over 12 years and changes in its social gradients in South Korea. PLoS ONE. 2018;13(9):e0204018. doi:10.1371/journal.pone.0204018
25. Bjerrum L, Søgaard J, Hallas J, Kragstrup J. Polypharmacy: correlations with sex, age and drug regimen. A prescription database study. Eur J Clin Pharmacol. 1998;54(3):197-202. doi:10.1007/s002280050445

26. Franchi C, Cartabia M, Risso P, et al. Geographical differences in the prevalence of chronic polypharmacy in older people: eleven years of the EPIFARM-Elderly Project. Eur J Clin Pharmacol. 2013;69(7):1477-1483. doi:10.1007/s00228-013-1495-7
27. Grimmsmann T, Himmel W. Polypharmacy in primary care practices: an analysis using a large health insurance database. Pharmacoepidemiol Drug Saf. 2009;18(12):1206-1213. doi:10.1002/pds.1841
28. Hovstadius B, Petersson G. The impact of increasing polypharmacy on prescribed drug expenditure—A register-based study in Sweden 2005–2009. Health Policy. 2013;109(2):166-174. doi:10.1016/j.healthpol.2012.09.005
29. Neutel CI, Skurtveit S, Berg C. Polypharmacy of potentially addictive medication in the older persons—quantifying usage. Pharmacoepidemiology and Drug Safety. 2012;21(2):199-206. doi:10.1002/pds.2214
30. Sørensen HT, Johnsen SP, Nørgård B. Methodological issues in using prescription and other databases in pharmacoepidemiology. Nor Epidemiol. 2001;11(1):13-18.
31. Bjerrum L, Rosholm JU, Hallas J, Kragstrup J. Methods for estimating the occurrence of polypharmacy by means of a prescription database. Eur J Clin Pharmacol. 1997;53(1):7-11. doi:10.1007/s002280050329
32. Colley CA, Lucas LM. Polypharmacy: the cure becomes the disease. J Gen Intern Med. 1993;8(5):278-283. doi:10.1007/BF02600099
33. Midão L, Giardini A, Menditto E, Kardas P, Costa E. Polypharmacy prevalence among older adults based on the survey of health, ageing and retirement in Europe. Arch Gerontol Geriatr. 2018;78:213-220. doi:10.1016/j.archger.2018.06.018
34. Castioni J, Marques-Vidal P, Abolhassani N, Vollenweider P, Waeber G. Prevalence and determinants of polypharmacy in Switzerland: data from the CoLaus study. BMC Health Serv Res. 2017;17. doi:10.1186/s12913-017-2793-z
35. Moen J, Antonov K, Larsson CA, et al. Factors associated with multiple medication use in different age groups. Ann Pharmacother. 2009;43(12):1978-1985. doi:10.1345/aph.1M354
36. Pappa E, Kontodimopoulos N, Papadopoulos AA, Tountas Y, Niakas D. Prescribed-drug utilization and polypharmacy in a general population in Greece: association with sociodemographic, health needs, health-services utilization, and lifestyle factors. Eur J Clin Pharmacol. 2011;67(2):185-192. doi:10.1007/s00228-010-0940-0
37. Hafferty JD, Campbell AI, Navrady LB, et al. Self-reported medication use validated through record linkage to national prescribing data. J Clin Epidemiol. 2018;94:132-142. doi:10.1016/j.jclinepi.2017.10.013

38. Nielsen MW, Søndergaard B, Kjøller M, Hansen EH. Agreement between self-reported data on medicine use and prescription records vary according to method of analysis and therapeutic group. J Clin Epidemiol. 2008;61(9):919-924. doi:10.1016/j.jclinepi.2007.10.021

39. Haider SI, Johnell K, Thorslund M, Fastbom J. Analysis of the association between polypharmacy and socioeconomic position among elderly aged > or =77 years in Sweden. Clin Ther. 2008;30(2):419-427. doi:10.1016/j.clinthera.2008.02.010

40. Jörgensen T, Johansson S, Kennerfalk A, Wallander M-A, Svärdsudd K. Prescription Drug Use, Diagnoses, and Healthcare Utilization among the Elderly. Ann Pharmacother. 2001;35(9):1004-1009. doi:10.1345/aph.10351

41. Onder G, Vetrano DL, Cherubini A, et al. Prescription Drug Use Among Older Adults in Italy: A Country-Wide Perspective. Journal of the American Medical Directors Association. 2014;15(7):531.e11-531.e15. doi:10.1016/j.jamda.2014.04.005

42. Cherubini A, Corsonello A, Lattanzio F. Polypharmacy in Nursing Home Residents: What Is the Way Forward? Journal of the American Medical Directors Association. 2016;17(1):4-6. doi:10.1016/j.jamda.2015.07.008

4.4. SUMMARY OF THE CHAPTER

In this chapter, the validity of BHIS-based information was tested across three topics in three different publications. The first article examined the validity of self-reported mammography uptake in a straightforward manner, using BCHI data as the gold standard. The second paper compared BHIS and BCHI data sources to ascertain the prevalence of a selection of CDs; while the third paper compared BHIS and BCHI data to estimate the prevalence of polypharmacy and assessed the complementarity of the two data sources.

In summary, the validity of BHIS-based information compared with information from the BCHI depends on the topics under consideration and is influenced by the characteristics of the respondents. There is a substantial over-estimation of BHIS based mammography uptake. This mis-estimation (overestimation) is mainly due to the telescoping. Therefore, BHIS data should not be used to estimate mammography screening coverage. To what concern the ascertainment of CDs, although the comparison of BHIS and BCHI data showed good agreement for some CDs such as diabetes, Parkinson's disease and thyroid disorders, a poor agreement was found for COPD and asthma. Next, estimating the prevalence of CDs from data on reimbursed drugs alone is not straightforward: caution should be exercised when using indicators based on these data alone to estimate the prevalence of CDs. The determinants of moderate polypharmacy do not vary significantly according to the source of the outcome, indicating that both BHIS and BCHI data sources are valid for estimating polypharmacy but the BHIS data are better suited to this purpose than the BCHI data because BCHI data do not include certain medicines, particularly non-prescribed or non-reimbursed medicines.

CHAPTER 5. USE OF LINKED DATA FOR LONGITUDINAL STUDY

PREDICTORS OF NURSING HOME ADMISSION IN THE OLDER POPULATION IN BELGIUM: A LONGITUDINAL FOLLOW-UP OF HEALTH INTERVIEW SURVEY PARTICIPANTS

The finding of this chapter were published as:

Berete F, Demarest S, Charafeddine R, De Ridder K, Vanoverloop J, Van Oyen H, Bruyère O and Van der Heyden J. Predictors of nursing home admission in the older population in Belgium: a longitudinal follow-up of health interview survey participants. BMC geriatrics 22.1 (2022): 1-13.

Berete *et al. BMC Geriatrics* (2022) 22:807
https://doi.org/10.1186/s12877-022-03496-4

BMC Geriatrics

RESEARCH Open Access

Predictors of nursing home admission in the older population in Belgium: a longitudinal follow-up of health interview survey participants

Finaba Berete[1,2*], Stefaan Demarest[1], Rana Charafeddine[1], Karin De Ridder[1], Johan Vanoverloop[3], Herman Van Oyen[1,4], Olivier Bruyère[5] and Johan Van der Heyden[1]

Abstract

Background: This study examines predictors of nursing home admission (NHA) in Belgium in order to contribute to a better planning of the future demand for nursing home (NH) services and health care resources.

Methods: Data derived from the Belgian 2013 health interview survey were linked at individual level with health insurance data (2012 tot 2018). Only community dwelling participants, aged ≥65 years at the time of the survey were included in this study ($n = 1930$). Participants were followed until NHA, death or end of study period, i.e., December 31, 2018. The risk of NHA was calculated using a competing risk analysis.

Results: Over the follow-up period (median 5.29 years), 226 individuals were admitted to a NH and 268 died without admission to a NH. The overall cumulative risk of NHA was 1.4, 5.7 and 13.1% at respectively 1 year, 3 years and end of follow-up period. After multivariable adjustment, higher age, low educational attainment, living alone and use of home care services were significantly associated with a higher risk of NHA. A number of need factors (e.g., history of falls, suffering from urinary incontinence, depression or Alzheimer's disease) were also significantly associated with a higher risk of NHA. On the contrary, being female, having multimorbidity and increased contacts with health care providers were significantly associated with a decreased risk of NHA. Perceived health and limitations were both significant determinants of NHA, but perceived health was an effect modifier on limitations and vice versa.

Conclusions: Our findings pinpoint important predictors of NHA in older adults, and offer possibilities of prevention to avoid or delay NHA for this population. Practical implications include prevention of falls, management of urinary incontinence at home and appropriate and timely management of limitations, depression and Alzheimer's disease. Focus should also be on people living alone to provide more timely contacts with health care providers. Further investigation of predictors of NHA should include contextual factors such as the availability of nursing-home beds, hospital beds, physicians and waiting lists for NHA.

Keywords: Nursing home admission, Institutionalization, Older adults, Predictors, Administrative data, Linkage, Competing risk analysis

5.1. ABSTRACT

Background

This study examines predictors of nursing home admission (NHA) in Belgium in order to contribute to a better planning of the future demand for nursing home (NH) services and health care resources.

Methods

Data derived from the Belgian 2013 health interview survey were linked at individual level with health insurance data (2012 tot 2018). Only community dwelling participants, aged ≥65 years at the time of the survey were included in this study (n=1930). Participants were followed until NHA, death or end of study period, i.e., December 31, 2018. The risk of NHA was calculated using a competing risk analysis.

Results

Over the follow-up period (median 5.29 years), 226 individuals were admitted to a NH and 268 died without admission to a NH. The overall cumulative risk of NHA was 1.4%, 5.7% and 13.1% at respectively 1 year, 3 years and end of follow-up period. After multivariable adjustment, higher age, low educational attainment, living alone and use of home care services were significantly associated with a higher risk of NHA. A number of need factors (e.g., history of falls, suffering from urinary incontinence, depression or Alzheimer's disease) were also significantly associated with a higher risk of NHA. On the contrary, being female, having multimorbidity and increased contacts with health care providers were significantly associated with a decreased risk of NHA. Perceived health and limitations were both significant determinants of NHA, but perceived health was an effect modifier on limitations and vice versa.

Conclusions

Our findings pinpoint important predictors of NHA in older adults, and offer possibilities of prevention to avoid or delay NHA for this population. Practical implications include prevention of falls, management of urinary incontinence at home and appropriate and timely management of limitations, depression and Alzheimer's disease. Focus should also be on people living alone to provide more timely contacts with health care providers. Further investigation of predictors of NHA should include contextual factors such as the availability of nursing-home beds, hospital beds, physicians and waiting lists for NHA.

Keywords: nursing home admission – institutionalization – older adults – predictors – administrative data – linkage – competing risk analysis

5.2. BACKGROUND

The ageing of populations combined with the increase in the prevalence of chronic conditions and the rapid advance in medical technology may lead to an increase of long-term care (LTC) services for older people [1]. In most industrialized countries, the demand for nursing homes (NH) is expected to rise sharply during the coming years [1–4]. According to the demographic projections of the Belgian Federal Planning Bureau in 2019, the share of the population aged 67 and over, which was 16% in 2018, will rise to 20% in 2030 and 23% in 2070 [5]. Therefore, the number of people requiring care either at home or at a NH will increase. Data from the Intermutualistic Agency (IMA) show that in 2020, 5,3% of the population aged 65 years and older resided in a NH [6]. If health policies focus on healthy ageing and the organization of well-functioning and integrated home care, the need for more NH could be reduced. According to a study conducted by the King Baudouin Foundation, 80% of Belgian older people wish to live at home as long as possible [7]. However, this can only be achieved in case of a suitable family or financial situation, or in the absence of serious medical problems. Otherwise, older people can rely upon a whole chain of care services, from home care through intermediate forms to permanent care. An intermediate form of services consists of day care centers and short-stay centers, while permanent care and support in a residential center is found at the end of the care services chain [8]. Day-care centers are a solution for people who are able to live at home but who do not have someone coming in daily to provide the necessary help and care. A short-stay center offers much of the same care as a permanent center for residential care, but the stay is limited in time (a maximum of 90 days, of which a maximum of 60 days can be consecutive, per year). In most cases short-stay centers are embedded in residential care centers, which then include a number of beds specifically designated for short-stay residents. As to the residential care centers in which older people stay permanently, a distinction should be made between a rest home, on the one hand, and a rest and care home, on the other hand. Only the latter are able to accommodate people with a high level of dependency and who require more extensive care. Most residential homes in Belgium are recognized as being rest

and care homes, which means that they include both residents with and without special care needs. To ease reading, both the rest homes and the rest and care homes will be referred to as nursing homes in this paper.

Although living independently at one's own home within the community is a major objective defining healthy ageing [9], the organization of LTC must take into consideration the balance between community and institutional care, which both have financial costs and societal impacts [9–11]. For a better planning of the future demand for NH services and health care resources, policy makers need to be aware of the predictors of nursing home admission (NHA). Previous studies have identified potential predictors of NHA on the basis of Andersen's behavioral model of health services use, which considers the use of health services to be a function of an individual's predisposing, enabling, and need characteristics [2,3,12–15]. Predisposing factors include demographics, social structure, and health beliefs. Enabling factors are those influencing an individual's ability to gain access to health services and include family and community resources. Need factors refer to the functional and health problems that generate the need for health care services [12]. The most relevant are marital status (being single or widowed), higher severity of cognitive impairment and mobility impairment [1], dementia [16,17], living situations and older age [2,10,12,18–20]. The role of urinary incontinence as predictor of NHA remains controversial. Some studies found that urinary incontinence is a strong predictor of NHA [21,22], while another study found that urinary incontinence was not an independent predictor of NHA after adjusting for confounders [23].

An important limitation of most previous studies on this topic is the lack of generalizability of the results because they are often conducted among specific subgroups such as patients with dementia or Alzheimer [1,16,17,24], myocardial infarction [19], or surgery as a result of a hip fracture [25]. Moreover, some of these studies lack power and precision since they are based on small samples [10]. A major methodological shortcoming in some studies is that they fail to take into account death as a competing risk of NHA in the analysis, which may bias the results [1,12,18,26]. Therefore, the objective of this study is to identify predictors of NHA in a Belgian community dwelling population aged 65 years and over [27], considering death as a competing risk factor.

5.3. METHODS

Study population and data

The data for this study were derived from a linkage at the individual level between data from the Belgian health interview survey (BHIS) of 2013 and data from the Belgian compulsory health insurance (BCHI) between 2012 and 2018. This linked data is further referred to as HISlink. The study population is limited to those aged 65 years and older.

The BHIS is a national, cross-sectional household survey conducted more or less every 5 years since 1997 by Sciensano, the Belgian institute for health. Participants are selected from the national population register through a multistage stratified sampling procedure. The participation rate was 57% at household level for the BHIS 2013. In the BHIS, information is collected on the health status, health behavior, health care consumption, use of medicines and sociodemographic characteristics. Post stratification weights were used to obtain representative results at the level of the Belgian population The detailed methodology of the survey is described elsewhere [27].

The BCHI data contain exhaustive and detailed information on the reimbursed health care of over 99% of the total population. These data were provided by IMA, a joint venture of the seven national health insurance organisations that collects and manages all data on healthcare expenditures. The BCHI contains three kinds of data: population data (a limited amount of demographic and socio-economic information), health care expenditure data (information on reimbursed health care) and pharmanet data (detailed information on all prescriptions for reimbursed drugs dispensed in public pharmacies) [6]. Although healthcare consumption is registered in detail, diagnostic information is not available. A proxy for diagnostic information over a number of chronic health conditions (e.g. cardiovascular disorder, diabetes, asthma, epilepsy, chronic obstructive pulmonary disease, Alzheimer's disease, etc.) is estimated through the volume of the prescription of reimbursed medication using an algorithm defined by a group of experts from the National Institute for Health and Disability Insurance (NIHDI). The algorithm is based on the anatomical, therapeutic, chemical (ATC) codes of specific drugs prescribed and dispensed in public pharmacies. A minimum threshold of 90 DDD (Defined Daily Dose) per year is used. For some

chronic conditions, the algorithm takes into account the age of the person to assign the disease or not. For instance, the proxy diagnosis of asthma is more likely to be attributed to people aged 50 years or younger, while those of chronic obstructive pulmonary disease is more likely attributed to people over 50 years to determine cases [28].

Individual BHIS 2013 data were linked with BCHI data using the unique national register number. The linkage rate was 96% for individuals aged 65 years and over. The BHIS sample includes both community dwelling and institutionalized people but for this study we only considered people aged 65 years and more residing at home at the moment of the interview (n = 1930).

Measures

Dependent variable

The main outcome was defined as a NHA after participating to the BHIS 2013 at any time during the follow-up period. The BHIS 2013 participants were followed until the date of NHA, the date of death or the end of study period, i.e., December 31, 2018. The information on the date of death was retrieved from the BCHI data. NHA was ascertained if at least one specific nomenclature code defined by the NIHDI for reimbursement purpose and corresponding to care delivery into home for older people or nursing home was found. The date of the first care delivery into NH was considered to be the date of admission into the NH. Only the first admission into NH was considered in this study. Short-stay care episodes, i.e. a maximum stay of 90 days, of which 60 days may be consecutive (defined by specific nomenclature codes) were excluded from the analyses. Therefore, this study focuses on NHA as a permanent resident. NHA for a short stay is not considered.

Independent variables

Independent variables were classified according to Andersen's behavioral model of health care use, as

1. predisposing variables: age, gender, educational attainment, living situation;
2. enabling factors: household income, appreciation of social contacts, home care services use, urbanization level and region of residence;

3. need factors: perceived health, Global Activity Limitations Indicator (GALI), multimorbidity, falls, urinary incontinence or problems in controlling the bladder in the past 12 months (either reimbursement for incontinence protections in BCHI data or self-reported), depression (self-reported depression in the past 12 months or self-reported use of anti-depressants in the past 2 weeks), Alzheimer's' disease, number of contacts with health care providers in the past 12 months (general practitioners, specialists, dentists, physiotherapists) and hospitalization in the past 12 months .

Home care services use is based on a single question: "In the last 12 months, have you received help at home or made use of home care services for yourself? (Yes / No)".This question is preceded by the following intro: "The next question is about home care services that cover a wide range of health and social services provided to people with health problems at their homes. These services comprise for example home care provided by a nurse or midwife, domestic help for older people, "meals on wheels" or transport service". Perceived health is based on the single question: "How is your health in general?". This question is part of the Minimum European Health Module (MEHM), which is internationally used. Five response categories are possible: Very good / Good / Fair / Poor / Very poor. The response categories Very good / Good are recorded as "Good to very Good" and those Fair / Poor / Very poor as "Very bad to fair".

As multimorbidity indicator we used the number of self-reported chronic conditions per person (out of a total of 25 chronic conditions), in the past 12 months. The list of these 25 chronic conditions is found in Table A1 (supplementary data). The GALI is an indicator of limitations due to health problems taking into account the person's environment and support. It is based on a single question asking the respondent to estimate the possible restrictions due to their health: "Have you been limited for at least 6 months because of a health problem in the activities that people usually do" (Yes, severely limited / Yes, limited / No, not limited at all) [29]. Information on contact with health care providers and hospitalization were obtained from the BCHI source. Urinary incontinence and Alzheimer's disease are combined indicators from BCHI and BHIS sources, while the other predictors were based on self-reported information. Alzheimer's disease cases were ascertained using the aforementioned experts' algorithm or the use of proxy interview because of a memory problem. The algorithm

is based on ATC codes of specific drugs (N06DX01, N06DA) prescribed and dispensed in public pharmacies. The minimum threshold of 90 DDD per year is used to determine cases [28]. Thus, Alzheimer's disease cases were identified as follows: "use of a minimum of 90 DDD per year of prescribed specific drugs (ATC codes = N06DX01, N06DA)" OR "having had a proxy interview because of a memory problem (e.g. amnesia, senile dementia). Detailed information about the variables description or operationalization are found in Table A1 (supplementary data).

Table 5.1 provides an overview of the independent variables (levels of variables, data sources and proportion of missing values, if any, before the multiple imputation).

Table 5.1: Overview of the covariates, according to the risk factor groups and data sources, HISlink 2013, Belgium

Covariates	**Level of covariates**	**Data sources**		**Missing values n (%)**
		BHIS	**BCHI**	
Predisposing				
Age		x		-
Gender	Male Female	x		-
Education	Low Middle High	x		19 (0.98)
Living situations	Live alone Not live alone	x		-
Enabling				
Household income	Low High	x		280 (14.5)
Level of urbanization	Urban Sub-urban Rural	x		-
Region of residence	Flanders Brussels Wallonia	x		-
Appreciation of social contacts	Rather unsatisfied Rather satisfied	x		440 (22.8)
Home care service use in the past 12 months preceding the survey	Yes No	x		3 (0.16)
Need				
Perceived health	Good to very good Very bad to fair	x		426 (22.1)

Multimorbidity		x		18 (0.88)
Long term limitation (GALI)	Yes, severely Yes No	x		452 (23.4)
Falls	Yes No	x		114 (5.91)
Urinary incontinence	Yes No	x	x	5 (0.26)
Depression	Yes No	x		8 (0.41)
Alzheimer's disease	Yes No	x	x	-
Number of contact with care providers in the past 12 months			x	-
Hospitalization in the past 12 months preceding the survey	Yes No		x	-

*HISlink = linkage between BHIS 2013 data and BCHI data from 2012 to 2018; * Missing values before the multiple imputation; GALI = Global Activity Limitation Indicator*

Statistical analyses

Descriptive statistics

Baseline characteristics described above were compared by NHA status (admitted to NH or not admitted to NH) with $\chi 2$ test for categorical variables, t-test for normally distributed continuous variables and Wilcoxon-Mann-Whitney test for non-normally distributed continuous variables.

Time to NHA was measured in years from the baseline survey to either the date of admission to a NH, date of death (without prior NHA) or end of study period, i.e., December 31, 2018 [20]. Participants who ended their follow-up period were censored [30]. The median follow-up time and median time to NHA and their interquartile range (IQR) were calculated.

Competing risk analysis

Competing risks occur frequently in the analysis of survival data [31]. A competing risk is an event whose occurrence precludes the occurrence of the primary event of interest [31–34]. For example, in a study examining time to death attributable to cardiovascular causes, death attributable to non-cardiovascular causes is a competing risk, because subjects who die from another cause are no longer at risk of death due to cardiovascular causes [31,32,35].In the same way, a study of time-to-NHA, death that occurs before NHA is a competing risk as it precludes NHA [36].

In studies involving older people, competing risk of death is especially high due to the higher mortality in this group. Therefore, there is a concern to account for participants who die without experiencing the study outcome of interest. Traditional approaches in survival analysis such as Kaplan-Meier survival analysis and Cox proportional hazards regression are not designed to take into account the competing risk of death [37] and will result in an overestimation of the effect [31,37]. The higher the death rate among the study population, the more substantial the overestimation.

To account for the presence of competing risks, Fine and Gray [38] proposed to apply the proportional subdistribution hazards model. In this model, estimates are based on modified hazards sets, where subjects experiencing the competing event are retained even after their competing event [33], unlike the Cox model. Whereas the cause-specific hazard function of the Cox model for an event of interest is the instantaneous rate of occurrence of that event in subjects who are currently event-free, the Fine and Gray subdistribution hazard function for a given event is the instantaneous rate of occurrence of that event in subjects who are either currently event-free or who have already experienced a concurrent event [31,35].

Studies of older individuals in which a substantial number of participants die during follow-up should use the competing risk analysis to accurately determine incidence and effect estimates [36,37].

Hence, in the current study, a competing risk regression model was used to estimate the association between predictors and NHA, treating death during follow-up as a competing risk.

The cumulative incidence function was used to estimate the risk of NHA over time (using Gray's method with death as a competing risk [36,39–41]) for the whole group. Furthermore, as GALI is one of the key indicators in this study due to the impact of limitations on NHA, we also stratified the cumulative incidence function by GALI with the Cuminc R-function [42]. Gray's test was used to examine for differences between the GALI strata [36,43].

In addition, we calculated sub-hazard ratios (sHR) of participants' characteristics for the risk of NHA, with death as a competing risk, also using the Fine and Gray's proportional sub-hazard model [30,33,36,38,39,43].

We imputed missing variables using the fully conditional specification method. The proportion of missing values ranged from 0.16% to 23% (Table 5.1). As all the variables with missing values are either binary or ordinary, a logistic regression method was used to impute missing values [30,44]. We created 20 imputed data sets. This number was large enough to achieve a very good efficiency (the relative efficiency was close to 1.0 for all effects). Covariates which were significantly ($P<0.05$) associated in the univariate analysis (Table A2 in Supplementary data) were retained in the final multivariable model [36]. Several interactions were tested between GALI and the other predictors (age, gender, perceived health and multimorbidity) to account for the possibility that the effect of limitations on the risk of NHA may depend on the level of other predictors. The HAZARDRATIO statement was used in the proportional hazards regression procedure in SAS (PROC PHREG) to produce custom hazard ratios for interactions [45].

For the sensitivity analyses, we used the Cox proportional model on both the imputed and non-imputed data (complete case analysis, n=1209). In the Cox proportional model, participants who died before ever being admitted into a NH and those who ended the study were censored. The results of the sensitivity analyses are reported in the supplementary data.

The assumption of proportional subdistribution hazards was evaluated by including interaction terms between the covariates and time [33,39]. The assumption was found to be met for all the covariates.

All statistical tests were 2-tailed and we used $p < 0.05$ to determine statistical significance. Analyses were performed using SAS® 9.4 taking into account the complex BHIS design. For the calculation of cumulative incidence function curves we used the package cmprsk by Gray [46] of the R project (R version 3.5.2 (2018-12-20).

5.4. RESULTS

Sample characteristics

The median follow-up time was 5.29 years (95% CI, 5.24-5.32), interquartile range (IQR), 5.12-5.59 years. At the end of follow-up, 226 individuals (13.3%) were admitted to NH with a median time to NHA of 3.0 years, (IQR: 1.5-4.6) and 268 individuals (13.2%) had died before potential NHA. Further information about the sample description (absolute incidences of NHA and death at 1-year follow-up, 3-year follow-up and at the end of study period) can be found in Figure A1 (supplementary data). Compared to those not admitted to a NH, participants who had been admitted to a NH were at baseline older, female, lower educated, and belonged more often to a lower income household. They also lived more often alone, were rather unsatisfied with their social contacts, had more often experienced falls in the year prior to the survey, and had more often health problems (bad perceived health, limitations, urinary incontinence, cognitive disorders (depression or Alzheimer's disease), but had less often multimorbidity. Further information on participants' characteristics can be found in Table 5.2.

Table 5.2: Baseline characteristics of the study population (n= 1930), HISlink 2013, Belgium (weighted)

	Admitted to nursing home N = 226 (13.3%)	**Not admitted to nursing home N = 1704 (86.7%)**	**P-value**
Age, mean (SD)	80.1 (0.59)	73.9 (0.26)	*<.0001*
Gender, n (%)			*0.0192*
Male	106 (45.8)	992 (57.6)	
Female	120 (54.2)	712 (42.4)	
Education, n (%)			*< 0.0001*
Low	144 (68.7)	713 (43.4)	
Middle	46 (17.0)	464 (28.8)	
High	36 (14.3)	527 (27.8)	
Household income, n(%)			*< 0.0001*
Low	180 (82.5)	1106 (66.6)	
High	46 (17.5)	598 (33.4)	
Level of urbanization, n (%)			*0.5857*
Urban	96 (36.9)	823 (42.4)	
Sub-urban	60 (34.2)	405 (30.6)	
Rural	70 (29.0)	476 (27.0)	
Region of residence, n (%)			*0.3470*
Flanders	83 (62.8)	630 (60.7)	
Brussels	37 (5.0)	363 (8.1)	
Wallonia	106 (32.1)	711 (31.2)	
Living situations, n (%)			*<.0001*
Live alone	130 (51.9)	496 (29.0)	
Not live alone	96 (48.1)	1208 (71.0)	
Appreciation of social contacts, n (%)			*0.0175*
Rather unsatisfied	32 (14.7)	132 (8.7)	
Rather satisfied	194 (85.3)	1572 (91.3)	
Home care service use in the past 12 months preceding the survey, n (%)			*<.0001*
Yes	102 (45.3)	346 (19.6)	
No	124 (54.7)	1358 (80.4)	
Perceived health, n (%)			*0.0049*
Good to very good	111 (50.4)	1113 (66.8)	
Very bad to fair	115 (49.6)	591 (33.2)	
Multimorbidity, median (Q1-Q3)*	1.64 (0.5-3.5)	1.79 (0.6-2.9)	*0.0067*
Long term limitation (GALI), n (%)			*0.0088*
Yes, severely	30 (16.1)	160 (10.2)	
Yes	87 (32.9)	477 (25.5)	
No	99 (51.0)	1067 (64.3)	
Falls, n (%)			*<.0001*
Yes	79 (36.2)	286 (16.9)	
No	147 (63.8)	1418 (83.1)	
Urinary incontinence, n (%)			*<.0001*
Yes	49 (22.8)	176 (10.7)	
No	177 (77.2)	1528 (89.3)	
Depression, n (%)			*<.0001*
Yes	50 (24.7)	195 (12.1)	
No	176 (75.3)	1509 (87.9)	
Alzheimer's disease, n (%)			*<.0001*
Yes	13 (12.6)	35 (1.7)	
No	213 (87.4)	1669 (98.3)	
Number of contact with care providers in the past 12 months, median (Q1-Q3)*	12.66 (7.9-18.7)	11.2 (6.3-17.7)	*0.0002*
Hospitalization in the past 12 months preceding the survey, n (%)			*0.0380*
Yes	59 (28.2)	323 (19.4)	
No	167 (71.8)	1381 (80.6)	

*HISlinkk = linkage between BHIS 2013 data and BCHI data from 2012 to 2018; Q1 = lower quartile, Q3 = upper quartile; GALI = Global Activity Limitation Indicator; *Non-parametric test (Wilcoxon Two-Sample Test).*

Cumulative risk of nursing home admission

The overall crude cumulative risk of NHA was 1.4% (95% CI, 0.9-2.1) at 1 year follow-up, 5.7% (95% CI, 4.7-6.8) at 3 years follow-up and 13.1% (95% CI, 11.3-15.0) at the end of follow-up (5.9 years). Figure 5.1 provides the unadjusted cumulative incidence curves for NHA, stratified by GALI, with death as a competing risk. The crude risk of NHA was significantly different in function of the severity of limitations (GALI) at least in one category as compared to the other categories (Gray's test: $\chi 2=22.28$, $p<0.0001$). Individuals who were severely limited had a higher risk of NHA at any time point of follow-up than those who were not. For instance, at the end of the study, the risk of NHA was significantly higher among individuals with severe limitations (20.0% (95% CI, 12.7-28.5)) than those without limitations (8.0% (95% CI, 6.2-10.1)). Whereas the risk of NHA was not statistically different between individuals with severe limitations (20.0% (95% CI, 12.7-28.5) as compared to those with moderate limitations (16.2% (95% CI, 11.6-21.5)), (Figure 5.1).

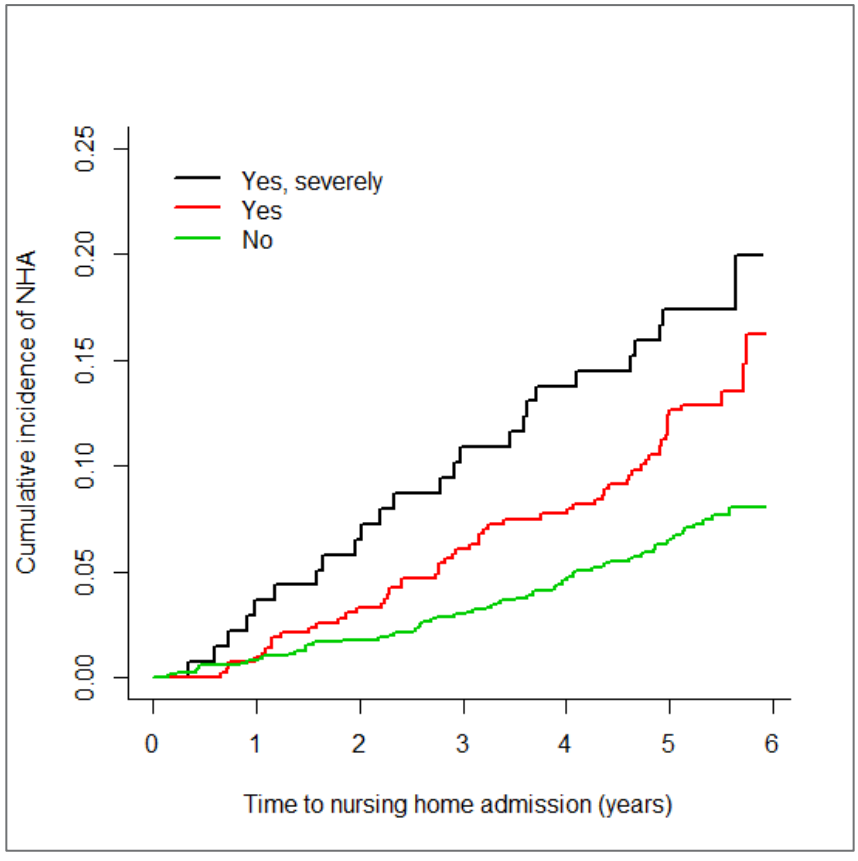


Figure 5.1: Crude cumulative risk of nursing home admission, stratified by limitations (GALI), accounting death as a competing risk, HISlink 2013, Belgium

Results of the subdistribution hazard model for nursing home admission

The results of the multivariable competing risk analysis are displayed in Table 5.3.

Among the predisposing factors, a one year increase in age is associated with a 9% increase risk of NHA. Individuals with lower educational attainment also showed a higher risk of NHA (sHR=1.44, 95% CI, 1.26-1.65) as compared to those with higher educational attainment, as well as living alone (sHR=1.68, 95% CI, 1.57-1.80). Being female, having an intermediate educational attainment were associated with a 25% and 14% reduction in the risk of NHA respectively.

Regarding the enabling factors, use of home care services in the past 12 months (sHR=1.57, 95% CI, 1.48-1.68) were associated with a higher risk of NHA.

With respect to the need factors, the interaction between limitations and perceived health was found to be statistically significant and interaction terms were therefore added in the final model. Because of this interaction, the interpretation of the sHRs of perceived health must take into account the levels of limitations and vice versa. Our results show for the impact of limitations, that 1) if people are in good perceived health, severe limitations (sHR=2.42, 95% CI, 2.35-2.49) increase the risk on NHA; and 2) if people are in bad perceived health, severe limitations (sHR=0.41, 95% CI, 0.40-0.42) decrease the risk of NHA. For the impact of perceived health, 1) if people have no limitations, bad perceived health increases the risk of NHA substantially (sHR=1.89, 95% CI, 1.85-1.92); and 2) if people have severe limitations, bad perceived health decreases the risk of NHA (sHR=0.30, 95% CI, 0.29-0.31). Further information about the results on interactions can be found in Table 5.3.

Individuals experiencing falls in the past 12 months (sHR=1.76, 95% CI, 1.64-1.89), those suffering from urinary incontinence (sHR=1.48, 95% CI, 1.22-1.79), and those suffering from depression (sHR=1.45, 95% CI, 1.25-1.70) had a higher risk of NHA than their counterparts. A one unit increase in the mean number of chronic conditions (multimorbidity) resulted in a 6% smaller risk of NHA (sHR=0.94, 95% CI, 0.90-0.97). An increased number of contacts with health care providers was associated with a decreased risk of NHA. Suffering from Alzheimer is the strongest predictor of NHA, with a sHR of 3.47 (95% CI, 3.05-3.95).

Table 5.3 : Predictors for nursing home admission: results from the competing risk regression (N=1930), HISlink 2013, Belgium

Potential predictors	**sHR (95% CI)**
Predisposing	
Age	1.09 (1.08-1.10)**
Female	0.75 (0.70-0.79)**
Educational attainment (Ref. = High)	
Low	1.44 (1.26-1.65)**
Middle	0.86 (0.75-0.98)*
Living alone	1.68 (1.57-1.80)**
Enabling	
Low household income	1.22 (1.00-1.50)
Unsatisfied with the social contacts	1.31 (0.73-2.37)
Home care service use in the past 12 months preceding the survey (Ref. =No)	1.57 (1.48-1.68)**
Need	
Long term limitation (GALI) and perceived health interaction[a]	
Bad perceived health vs. good perceived health at severe limitations	0.30 (0.29-0.31)*
Bad perceived health vs. good perceived health at moderate limitations	0.97 (0.95-0.98)*
Bad perceived health vs. good perceived health at no limitations	1.89 (1.85-1.92)*
Severe limitations vs. no limitations at good perceived health	2.42 (2.35-2.49)*
Severe limitations vs. no limitations at bad perceived health	0.41 (0.40-0.42)*
Severe limitations vs. moderate limitations at good perceived health	2.09 (2.03-2.16)*
Severe limitations vs. moderate limitations at bad perceived health	0.65 (0.64-0.66)*
Moderate limitations vs. no limitations at good perceived health	1.22 (1.20-1.24)*
Moderate limitations vs. no limitations at bad perceived health	0.64 (0.63-0.66)*
Multimorbidity	0.94 (0.90-0.97)*
Falls (Ref. = No)	1.76 (1.64-1.89)**
Urinary incontinence (Ref. = No)	1.48 (1.22-1.79)*
Depression (Ref. = No)	1.45 (1.25-1.70)**
Alzheimer disease (Ref. = No)	3.47 (3.05-3.96)**
Number of contact with health care providers in the past 12 months preceding the survey	0.98 (0.97-0.99)**
Hospitalization in the past 12 months preceding the survey (Ref. = No)	1.07 (0.95-1.21)

*HISlink = linkage between BHIS 2013 data and BCHI data from 2012 to 2018; [a] The HAZARDRATIO statement was used in PROC PHREG to produce custom hazard ratios for interactions; NHA = nursing home admission; sHR = sub Hazard Ratios; GALI = Global Activity Limitation Indicator. To facilitate reading, the GALI categories were reported as severe limitations, moderate limitations and no limitations, which referred to the categories yes, severely limited, yes limited and no, not limitation at all respectively; * $p < 0.05$; ** $p < 0.0001$.*

5.5. DISCUSSION

Summary of the results

To our knowledge this is the first study that investigated predictors of NHA among the Belgian community dwelling population aged 65 years and older.

The overall unadjusted cumulative incidence (risk) of NHA, accounting for death as competing risk, was of 5.7% at 3 years of follow-up and of 13.1% at the end of the study. After adjusting for baseline characteristics of participants, important predictors of NHA were found. These were, among others, being older, living alone, having used of home care services, having a history of falls, depression or Alzheimer's disease, all of which were significantly associated with a higher risk of NHA. Predictors such as being female, having multimorbidity and increased contacts with health care providers were significantly associated with a decreased risk of NHA.

Incidence of NHA

The incidence of NHA may be influenced by organizational aspects and cultural aspects [47], but also by the availability, accessibility and affordability of home care facilities. The characteristics of the study population also play a role since the cumulative incidence of NHA could be affected by the higher risk of death in the population under study. Previous studies have investigated predictors of NHA among sub-groups of the population. For instance, Bergkamp et al. [36] investigated predictors of NHA in Cerebral Small Vessel Disease (CSVD) patients and have found that after 5-years follow-up, the cumulative incidence was 3.6% (95% CI, 2.2-5.5) and 6% (95% CI, 4.2–8.3) after 8 years of follow-up. This cumulative incidence is lower than those found in our study. A possible explanation could be that the risk of the competing event, i.e., death, is likely to be higher in patients suffering from CSVD than in the general population, which may affect the risk of NHA. In contrast, our cumulative incidence is lower than the cumulative incidence found by Wolff et al. in community living older adults in the USA (16.1% in a 2-years follow-up) [48]. This difference could be explained by differences in the characteristics of the study population. Indeed, the Wolff et al. study participants were much older (sample mean age of 79 years compared to 74.7 years in our study), received assistance with personal care or mobility from a family member or unpaid caregiver (help with 2 of 6 self-care activities) and nearly 1 in 3 had dementia.

Factors associated with NHA

Important predictors of NHA were identified in our study. This will be discussed in the following paragraphs in relation to findings from previous studies and according to the Andersen behavioural model.

In earlier studies, advanced age emerged as a strong predictor of NHA among the predisposing factors [12,25,26]. In accordance with these studies, higher age was also found to be a significant predictor of NHA, even after taking into account the competing risk of death. With regard to gender, women were found to be less likely to enter NH than men. This result is consistent with the study by Gaugler et al. [26] (female: HR of risk: 0.87 (95% CI, 0.81-0.93)). Although this is an unexpected result compared to the results presented in Table 5.2 and the univariate analysis (Table A2, supplementary data), it could be explained by a protective effect of the female sex against NHA compared to men. Previous studies have also found a protective effect of female gender against NHA [25,26,49]. Casanova et al (2021) argue that the protective effect of female gender against NHA may be explained by a stronger negative preference for NH care among women or by the fact that children provide more informal care for women than for men [49]. Living situations appeared as the strongest predictor of NHA among the predisposing factors. Individuals who lived alone had nearly twice the sHR to enter a NH. Our findings are in line with those in previous studies [20,50]. Although older people prefer living in their own home as long as possible [7,11,51] this may be more difficult for people living alone because of lack of social support and lack of informal care.

Among the enabling factors, the use of home care services in the previous year was associated with a greater risk of NHA. This result is not surprising since the use of home care services is generally an expression of a need for support and therefore a first step towards possible NHA. Our result is similar to those in a study on predictors of NHA after hip fracture. The authors found that receiving home care before injury was associated with an increase in HR of 2.00 (95% CI 1.54-2.61), HR 1.64 (95% CI, 1.43-1.87), and HR 1.22 [95% CI, 1.13-1.32) for patients aged 60 to 69 years, 70 to 79 years, and 80 to 89 years respectively [25].

Within the need factors, if either poor perceived health or severe limitations are present there is an increased risk of NHA, but when they occur together the risk of NHA decreases, most likely because for those people the risk of dying is larger than

the risk of being admitted to a NH (competing risk). For instance, the risk of dying in case of bad perceived health and severe limitations at 1, 3 and 5-years follow-up is 8.5%, 23.9% and 35% respectively. The risk of NHA in case of bad perceived health and severe limitations at the same time points is 2.6%, 9.4% and 16.5% respectively (Table A5, supplementary data). The paradoxical finding of a decreasing risk of NHA when both poor perceived health and severe limitations are present is in line with the fact that an increasing number of chronic diseases was associated with a reduced risk of NHA, probably because of competing risk of death among this group. Indeed, the higher the number of chronic conditions the higher the risk of poor perceived health and more severe limitations.

In line with earlier studies [26,52], we found that a history of falls in the past 12 months was associated with an increased risk of NHA. In fact, in some cases, falls among the older people can lead to more serious events (fractures, injuries, loss of autonomy) with adverse consequences on their health status and therefore precipitate their admission to a NH.

The presence of Alzheimer's disease is by far the strongest predictor of NHA. In the literature, beside age, cognitive comorbidities (depression, Parkinson, dementia or Alzheimer's disease) and functional impairment were among the strongest predictors and are associated with an increased risk of NHA [12,15,25,26,53]. For example, in a study among a general older population, the authors found that Alzheimer's or dementia increases the hazard of NHA by 20.2 times for men and 10.0 times for women [12]. In another study of 137,000 community dwelling patients aged 65 years or more, Harris et al. found that depression was associated with a higher risk of NHA in the general population [53]. Another interesting but surprising result is that with an increasing number of contacts with health care providers, the risk of NHA decreased. This finding could be explained on the one hand by the fact that people with a higher number of contacts are likely to be in poorer health and therefore less likely to enter a nursing home due to a higher risk of death. On the other hand, an increased number of contacts with health care providers will allow appropriate treatment and therefore prevent or delay NHA. Luppa et al. (2010) [18] also found a decreased risk of NHA with an increased number of specialist visits and explained this finding as a positive effect of appropriate treatment of medical conditions by specialists. Other need factors are of lesser importance.

Strengths and limitations

From a public health perspective, the major strengths of this study include the use of a large sample and the use of a large number of individual-level predictors, a relatively long follow-up period, and the linkage to administrative data to identify NHA and/or death. The use of the competing risk analysis is another strength of this study. Indeed, we performed competing risk regression to study the association between several covariates and the risk of NHA. This approach is preferred over a standard survival model because in older population, death may compete with NHA, and ignoring such competing risk may lead to biased results [37]. In competing risk situations, the cumulative incidence function was more appropriate as it took competing events into account when estimating the incidence. We further chose the Fine and Gray model over the cause-specific hazard model as our primary interest was in predictive modelling.

The current study has some limitations that deserves to be pointed out. First, the exact dates of NHA were not available. We used the dates of the first registered care in a NH based on specific nomenclature code as a proxy of dates of NHA. However, these dates are quite accurate and deviations from the exact dates are small. Second, almost all covariates included in the analysis were measured at baseline and most of them are self-reported. Therefore, possible changes (e.g., in living situations or social support) over the course of the study are not taken into account and the risk of reporting bias remains. Third, although in recent years efforts have been made to avoid NHA by taking measures to facilitate home health care, our data did not allow demonstrating this. To investigate this thoroughly, longitudinal data are required on both the evolution of the health situation and the use of home health care, but unfortunately such information was not available in our study. Fourth, data on local variations in supply of care and/or home care services (supply of NH beds, hospital beds, and physicians in the region of residence, waiting lists, etc.) as potential important confounders at the enabling level were unavailable and therefore not included in our analyses. Fifth, due to the lack of diagnostic information, Alzheimer's disease indicator was estimated based on prescribed specific medications and self-reported information on memory problem (e.g. amnesia, senile dementia), making it less sensitive. Indeed, many people with Alzheimer do not take specific medications. So individuals suffering from this disease might be underestimated. Sixth, although

this study was conducted in a large dataset, a selection bias is unavoidable and the representativeness of the sample is not guaranteed. However, through the calculation of post stratification weights, with the Belgian national registry as auxiliary data source, results are as representative as possible of the Belgian community dwelling population. Finally, the finding of this study may not be generalized to other areas or settings with lower health system standards, for example because the organization of the Belgian health care system can be very different from other countries.

Implications and challenges for the future

This study has implications for practitioners and policy makers. As a result of the ageing population the pressure on NH will only increase. Efforts and measures that enable older people to remain longer at home will not only be beneficial from a budgetary point of view but will also increase the wellbeing of older people. This study identified some domains in which health care professionals and policy makers should further invest to delay NHA. Prevention of falls is a first important point. Home care givers should also be trained to deal with mental health problems. Attention should also be paid to the problems of urinary incontinence of older people. Adaptation of the home environment in a way that despite their limitations, older people can still continue their daily activities is of utmost importance. Finally, the strong association found between Alzheimer's disease and NHA is of course not surprising. Alzheimer's disease, dementia and severe cognitive problems are important reasons why people have to be admitted in a NH. Therefore, at population level, further efforts are needed to prevent important risk factors for dementia. Focus should also be on people living alone to provide the appropriate social support and more timely contacts with health care providers. Further investigation of predictors of NHA should include contextual enabling actors such as the supply of nursing-home beds, hospital beds, physicians and waiting lists for NHA. Analysis taking into account other competing events such as home health care services should also be considered.

5.6. CONCLUSIONS

Our findings underline important predictors of NHA of older adults, and therefore offer possibilities of prevention to avoid or delay NHA for this population. Practical implications include prevention of falls, management of people with urinary incontinence at home, and appropriate and timely management of limitations, depression and Alzheimer's disease. Focus should also be on people living alone to provide more timely contacts with health care providers. Further investigation of predictors of NHA should include contextual enabling factors such as the supply of nursing-home beds, hospital beds, physicians and waiting lists for NHA.

5.7. BIBLIOGRAPHY

1. Luppa M, Riedel-Heller SG, Stein J, Leicht H, König H-H, van den Bussche H, et al. Predictors of Institutionalisation in Incident Dementia – Results of the German Study on Ageing, Cognition and Dementia in Primary Care Patients (AgeCoDe Study). Dement Geriatr Cogn Disord. 2012;33:282–8.
2. Stolz E, Mayerl H, Rásky É, Freidl W. Individual and country-level determinants of nursing home admission in the last year of life in Europe. Vellakkal S, editor. PLoS ONE. 2019;14:e0213787.
3. Friedman SM, Steinwachs DM, Rathouz PJ, Burton LC, Mukamel DB. Characteristics Predicting Nursing Home Admission in the Program of All-Inclusive Care for Elderly People. The Gerontologist. 2005;45:157–66.
4. Van den Bosch K, Willemé P, Geerts J, Breda J, Peeters S, Van De Sande S, et al. Soin résidentiels pour les personnes âgées en Belgique : projections 2011-2025. Health Services Research (HSR). Bruxelles: Centre fédéral d'expertise des soins de santé (KCE); 2011. Report No.: KCE Reports 167B. D/2011/10.273/64.
5. Bureau fédéral du Plan et Statbel. Perspectives démographiques 2018_2070. Population et ménages [Internet]. 2019 Jan. Report No.: D/2019/7433/1. Available from: https://www.plan.be/uploaded/documents/201901240958590.FOR_POP1870_11813_F.pdf
6. Agence InterMutualiste -InterMutualistisch Agentschap (AIM-IMA). Agence InterMutualiste -InterMutualistisch Agentschap [Internet]. [cited 2021 Jul 26]. Available from: https://www.ima-aim.be/-Donnees-de-sante-
7. Fondation Roi Baudouin. Choix de Vie Durant Les Vieux Jours : Enquête Auprès de plus de 2000 Personnes de 60 Ans et + [Internet]. Brussels: Belgian; 2017. Available from: http://lampspw.wallonie.be/dgo4/site_colloques/ConceptionAdaptable/assets/documents/presentation/fondation-roi-baudouin-choix-vie-60-ans-et-plus.pdf
8. Structures d'hébergement et de soins [Internet]. Belgium.be Informations et services officiels. 2022. Available from: https://www.belgium.be/fr/sante/soins_de_sante/services_medicaux/maisons_de_repos
9. Nuutinen M, Leskelä R-L, Suojalehto E, Tirronen A, Komssi V. Development and validation of classifiers and variable subsets for predicting nursing home admission. BMC Med Inform Decis Mak. 2017;17:39.
10. Hajek A, Brettschneider C, Lange C, Posselt T, Wiese B, Steinmann S, et al. Longitudinal Predictors of Institutionalization in Old Age. Federici S, editor. PLoS ONE. 2015;10:e0144203.

11. Grenade L, Boldy D. Social isolation and loneliness among older people: issues and future challenges in community and residential settings. Aust Health Review. 2008;32:468.
12. Tomiak M, Berthelot J-M, Guimond E, Mustard CA. Factors Associated With Nursing-Home Entry for Elders in Manitoba, Canada. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences. 2000;55:M279–87.
13. Andersen RM. National Health Surveys and the Behavioral Model of Health Services Use. Medical Care. 2008;46:647–53.
14. Bell CL, LaCroix AZ, Desai M, Hedlin H, Rapp SR, Cene C, et al. Factors Associated with Nursing Home Admission after Stroke in Older Women. Journal of Stroke and Cerebrovascular Diseases. 2015;24:2329–37.
15. Luppa M, Luck T, Weyerer S, Konig H-H, Brahler E, Riedel-Heller SG. Prediction of institutionalization in the elderly. A systematic review. Age and Ageing. 2010;39:31–8.
16. Gaugler JE, Yu F, Krichbaum K, Wyman JF. Predictors of Nursing Home Admission for Persons with Dementia. Medical Care. 2009;47:191–8.
17. Luppa M, Luck T, Braumlhler E, K&oumlnig H-H, Riedel-Heller SG. Prediction of Institutionalisation in Dementia. Dement Geriatr Cogn Disord. 2008;26:65–78.
18. Luppa M, Luck T, Matschinger H, König H-H, Riedel-Heller SG. Predictors of nursing home admission of individuals without a dementia diagnosis before admission - results from the Leipzig Longitudinal Study of the Aged (LEILA 75+). BMC Health Serv Res. 2010;10:186.
19. Smedegaard L, Kragholm K, Numé A-K, Charlot MG, Gislason GH, Hansen PR. Nursing home admission after myocardial infarction in the elderly: A nationwide cohort study. Van Bogaert P, editor. PLoS ONE. 2018;13:e0202177.
20. Pimouguet C, Rizzuto D, Schön P, Shakersain B, Angleman S, Lagergren M, et al. Impact of living alone on institutionalization and mortality: a population-based longitudinal study. Eur J Public Health. 2016;26:182–7.
21. Ouslander JG, Kane RL, Abrass IB. Urinary Incontinence in Elderly Nursing Home Patients. JAMA. 1982;248:1194–8.
22. Thom DH, Haan MN, Van Den Eeden SK. Medically recognized urinary incontinence and risks of hospitalization, nursing home admission and mortality. Age Ageing. 1997;26:367–74.
23. Holroyd-Leduc JM, Mehta KM, Covinsky KE. Urinary Incontinence and Its Association with Death, Nursing Home Admission, and Functional Decline: URINARY INCONTINENCE AND ADVERSE OUTCOMES. Journal of the American Geriatrics Society. 2004;52:712–8.
24. Eska K, Graessel E, Donath C, Schwarzkopf L, Lauterberg J, Holle R. Predictors of Institutionalization of Dementia Patients in Mild and Moderate Stages: A 4-Year Prospective Analysis. Dement Geriatr Cogn Disord Extra. 2013;3:426–45.

25. Wahlsten LR, Smedegaard L, Brorson S, Gislason G, Palm H. Living settings and cognitive impairment are stronger predictors of nursing home admission after hip fracture surgery than physical comorbidities A nationwide Danish cohort study. Injury. 2020;51:2289–94.
26. Gaugler JE, Duval S, Anderson KA, Kane RL. Predicting nursing home admission in the U.S: a meta-analysis. BMC Geriatr. 2007;7:13.
27. Demarest S, Van der Heyden J, Charafeddine R, Drieskens S, Gisle L, Tafforeau J. Methodological basics and evolution of the Belgian health interview survey 1997–2008. Arch Public Health. 2013;71:24.
28. Echantillon Permanente Steekproef EPS R13 – FLAGS Release 20190201 FR [Internet]. 2019 [cited 2021 Jul 26]. Available from: https://aim-ima.be/IMG/pdf/eps_r13_-_flags_release_20190201_fr_-_vs2.pdf
29. Van Oyen H, Bogaert P, Yokota RTC, Berger N. Measuring disability: a systematic review of the validity and reliability of the Global Activity Limitations Indicator (GALI). Arch Public Health. 2018;76:25.
30. Brown RT, Diaz-Ramirez LG, Boscardin WJ, Lee SJ, Williams BA, Steinman MA. Association of Functional Impairment in Middle Age With Hospitalization, Nursing Home Admission, and Death. JAMA Intern Med. 2019;179:668.
31. Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation. 2016;133:601–9.
32. Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. Statistics in Medicine. 2017;36:4391–400.
33. Kohl M, Plischke M, Leffondré K, Heinze G. PSHREG: A SAS macro for proportional and nonproportional subdistribution hazards regression. Computer Methods and Programs in Biomedicine. 2015;118:218–33.
34. Guo C, So Y. Cause-Specific Analysis of Competing Risks Using the PHREG Procedure. SAS Global Forum. 2018;2018:18.
35. Lau B, Cole SR, Gange SJ. Competing Risk Regression Models for Epidemiologic Data. American Journal of Epidemiology. 2009;170:244–56.
36. Bergkamp MI, Wissink JGJ, van Leijsen EMC, Ghafoorian M, Norris DG, van Dijk EJ, et al. Risk of Nursing Home Admission in Cerebral Small Vessel Disease: Association With Lower Brain and White Matter Volumes. Stroke. 2018;49:2659–65.
37. Berry SD, Ngo L, Samelson EJ, Kiel DP. Competing Risk of Death: An Important Consideration in Studies of Older Adults: Competing risk of death in studies of older adults. Journal of the American Geriatrics Society. 2010;58:783–7.
38. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999;94:496–509.

39. Reber KC, Lindlbauer I, Schulz C, Rapp K, König H-H. Impact of morbidity on care need increase and mortality in nursing homes: a retrospective longitudinal study using administrative claims data. BMC Geriatr. 2020;20:439.
40. Gillam MH, Ryan P, Graves SE, Miller LN, de Steiger RN, Salter A. Competing risks survival analysis applied to data from the Australian Orthopaedic Association National Joint Replacement Registry. Acta Orthopaedica. 2010;81:548–55.
41. Gray RJ. A class of k-sample test for comparing the cumulative incidence of a competing risk. The Annals of Statistics. 1988;16:1141–54.
42. Brock G, Barnes C, Ramirez J, Myers J. R code for calculating the competing risks estimates. 2011;6.
43. Huyer G, Brown CRL, Spruin S, Hsu AT, Fisher S, Manuel DG, et al. Five-year risk of admission to long-term care home and death for older adults given a new diagnosis of dementia: a population-based retrospective cohort study. CMAJ. 2020;192:E422–30.
44. Berglund PA. Multiple Imputation Using the Fully Conditional Specification Method: A Comparison of SAS®, Stata, IVEware, and R. Proceedings of the SAS Global Forum 2015 Conference Cary, NC: SAS Institute Inc. 2015;17.
45. Savarese PT, Patetta MJ. 253-2010: An Overview of the CLASS, CONTRAST, and HAZARDRATIO Statements in the SAS® 9.2 PHREG Procedure [Internet]. 2010 [cited 2021 Dec 25]. Available from: resources/papers/proceedings10/253-2010.pdf
46. Gray B. Package 'cmprsk. Subdistribution analysis of competing risks. R package version 2, 2–10 [Internet]. 2020 [cited 2021 Jul 26]. Available from: http://cran.uvigo.es/web/packages/cmprsk/cmprsk.pdf
47. Cegri F, Orfila F, Abellana RM, Pastor-Valero M. The impact of frailty on admission to home care services and nursing homes: eight-year follow-up of a community-dwelling, older adult, Spanish cohort. BMC Geriatr. 2020;20:281.
48. Wolff JL, Mulcahy J, Roth DL, Cenzer IS, Kasper JD, Huang J, et al. Long-Term Nursing Home Entry: A Prognostic Model for Older Adults with a Family or Unpaid Caregiver. J Am Geriatr Soc. 2018;66:1887–94.
49. Casanova M. Revisiting the Role of Gender and Marital Status as Risk Factors for Nursing Home Entry. Ailshire J, editor. The Journals of Gerontology: Series B. 2021;76:S86–96.
50. Pynnonen K, Tormakangas T, Heikkinen R-L, Rantanen T, Lyyra T-M. Does Social Activity Decrease Risk for Institutionalization and Mortality in Older People? The Journals of Gerontology Series B: Psychological Sciences and Social Sciences. 2012;67:765–74.
51. Prieto-Flores M-E, Forjaz MJ, Fernandez-Mayoralas G, Rojo-Perez F, Martinez-Martin P. Factors Associated With Loneliness of Noninstitutionalized and Institutionalized Older Adults. J Aging Health. 2011;23:177–94.

52. Sørbye LW, Sørbye LW, Hamran, Henriksen, Norberg. Home care patients in four Nordic capitals – predictors of nursing home admission during one-year followup. JMDH. 2010;11.
53. Harris Y, Cooper JK. Depressive Symptoms in Older People Predict Nursing Home Admission: Depression predicts nursing home admission. Journal of the American Geriatrics Society. 2006;54:593–7.

CHAPTER 6. USE OF LINKED DATA TO ANSWER POLICY DRIVEN QUESTIONS - FURTHER ADDED VALUE

DOES HEALTH LITERACY MEDIATE THE RELATIONSHIP BETWEEN SOCIOECONOMIC STATUS AND HEALTH RELATED OUTCOMES IN THE BELGIAN ADULT POPULATION?

The findings of this chapter have been submitted as:

Berete F, Gisle L, Demarest S, Charafeddine R, Bruyère O, Van den Broucke S, Van der Heyden J. Does health literacy mediate the relationship between socioeconomic status and health related outcomes in the Belgian adult population? to BMC Public Health.

6.1. ABSTRACT

Background

Health literacy (HL) has been put forward as a potential mediator through which socioeconomic status (SES) affects health. This study explores whether HL mediates the relation between SES and a selection of health or health related outcomes.

Methods

Data from the participants of the Belgian health interview survey (BHIS) 2018 aged 18 years or older were individually linked with data from the Belgian compulsory health insurance (n=6878). HL was assessed with the HLS-EU-Q6. Mediation analysis were performed with health behaviour (physical activity, diet, alcohol and tobacco consumption), health status (perceived health status, mental health status), use of medicine (purchase of antibiotics), and use of preventive care (preventive dental care, influenza vaccination, breast cancer screening) as dependent outcome variables, educational attainment and income as independent variables of interest, age and sex as potential confounders and HL as mediating variable.

Results

The study showed that unhealthy behaviours (except alcohol consumption), poorer health status, higher use of medicine and lower use of preventive care (except flu vaccination) were associated with low SES (i.e., low education and low income) and with insufficient HL. HL partially mediated the relationship between education and health behaviour (except tobacco consumption), perceived health status and preventive dental care, accounting for 4.4% to 15.4% of the total effect. HL also constituted a pathway by which income influences health behaviour (except alcohol consumption), perceived health status, mental health status and preventive dental care, with the mediation effects accounting for 4.2% to 12.0% of the total effect.

Conclusions

Although the influence of HL in the pathway is limited, our findings suggest that strategies for improving various health related outcomes among low SES groups should include initiatives to enhance HL in these population groups. Further research is needed to confirm our results and to better explore the mediating effects of HL.

Keywords: health disparities; health literacy; mediation analysis; socioeconomic status.

6.2. INTRODUCTION

There is strong evidence that socioeconomic status (SES) is an important determinant of health disparities between population groups, with low SES being associated to poorer health conditions and less healthy behaviours (1–3). Several factors and mechanisms have been proposed to explain the chain of events linking SES to health outcomes (2), including material circumstances (like living and working conditions), behavioural factors, social cohesion and social capital and lack of social support, as well as psychological factors like stress, social comparison, less coping resources and skills. However, the pathway through which SES exerts its effect on health has not yet been fully clarified (4).

Health literacy (HL) has been hypothesized as a potential mediator through which SES affects health (5–10). According to the European Health Literacy Survey (HLS-EU) Consortium and the Health promotion glossary 2021, health literacy "is linked to literacy and entails a person's knowledge, motivation and competences to access, understand, appraise, and apply health information in order to make judgments and take decisions in everyday life concerning healthcare, disease prevention and health promotion to maintain or improve quality of life during the life course" (14,15). The mediating effect of HL is assumed to be especially important for behaviours for which individual judgement and decision making are necessary, such as physical activity and diet (11) or self-rated health status (8,9,12,13).

HL is an important factor for assessing public and personal health outcomes. A number of studies showed associations between low levels of HL and poorer health conditions (14,15), more frequent use of health services, longer hospitalisations (14,16) and higher mortality (15,17). Moreover, low level of HL has been associated with unhealthy behaviours, such as smoking (18,19), low physical activity (19,20) and less use of preventive services (15,18). On the other hand, HL has been associated with socioeconomic indicators such as educational attainment, income (9), material and social wealth or deprivation, unemployment status, occupation, as well as the sociodemographic profile (sex, age) of individuals (21). In view of this, the World Health Organisation considers HL as an important determinant of health, influenced by socioeconomic and cultural characteristics of the population, and by the degree of complexity of the health systems (22). As such, HL can be taken into account in efforts to reduce health disparities. Indeed, if HL is an important mediator in explaining

socioeconomic (SE) health differences, actions to improve HL in low SE groups will reduce disparities.

In Belgium, equity in the use of healthcare resources is an important concern. However, empirical research investigating the contribution of HL in the relationship between SES and health remains scarce. Furthermore, insight is needed concerning the link with other factors that play a role in health inequities. Health related outcomes that can be studied in this perspective, and for which data are available, include 1) health behaviours (physical activity, diet, alcohol and tobacco consumption), 2) health status (perceived health status, mental health), 3) use of medicine (purchase of antibiotics), and 4) use of preventive care (preventive dental care, influenza vaccination, breast cancer screening). These factors have been selected because a mediating effect of HL can be expected, given that each of them requires individual judgement and decision-making. More specifically, the hypothesis is that people with insufficient or limited HL have unhealthy behaviour, understand health promotion and intervention programmes less well and manage their health problems less well, resulting in poorer health status.

The availability of linked subjective and objective data (HISlink data) makes it possible to explore these different areas of interest and to test mediation hypotheses.

The purpose of the present study is to determine whether HL mediates the associations between education and income (SES) and the above mentioned health related outcomes. More specifically, the objectives are as follows:

1) to explore the association between SES and HL
2) to examine the association between SES and the selected health related outcomes
3) to examine he association between HL and the selected health related outcomes
4) to investigate the mediation effects of HL in the relationship between SES and the selected health related outcomes.

Educational attainment and income are both explored as independent variables as a previous study has shown that the relationship between HL and income is independent of educational attainment (23).

6.3. METHODS

Data and study population

The participants of this study were involved in the Belgian Health interview Survey (BHIS) 2018. Participants for the BHIS are selected through a stratified multistage clustered sampling design (24). The target population consists of all residents living in private households in Belgium and people who live in nursing homes. The BHIS collects information on the health status, health behaviour, HL, health care consumption, use of medicines and sociodemographic characteristics of all participants.

The BHIS data were individually linked to the Belgian Compulsory Health Insurance (BCHI) data using the unique national register number (HISlink 2018). The BCHI data contain exhaustive and detailed information on the reimbursed health expenses of over 99% of the total population. The database also includes a limited amount of sociodemographic information. The BCHI data were provided by the Intermutualistic Agency (IMA). IMA is a joint venture of the seven national health funds and collects and manages all data on healthcare expenditures as well as prescription information on reimbursed medicines (Pharmanet data) (25). Pharmanet records all data on reimbursed medication dispensed from public pharmacies in Belgium. Pharmanet data include information on the date of dispensing, the quantity per package, the daily defined dose and the national code number of the medicine which allows to link each medicine to its ATC-code.

Of the total of 11611 individuals who participated in the BHIS 2018, 10933 had their data linked with BCHI data, resulting in an overall linkage rate of 94%. In the BHIS, questions on HL were only addressed to people aged 15 years and over, in the form of self-report. Because younger individuals may be dependent of their parents' lifestyle and literacy in health and because the HL instrument was validated for people aged 18 years and over, this study is limited to adults aged 18 years or more (n=6682), except for breast cancer screening (recommended for women aged 50-69 years) and flu vaccination (recommended for the 65 years or older). Proxy interviews as well as missing HL records were excluded, Figure 6.1.

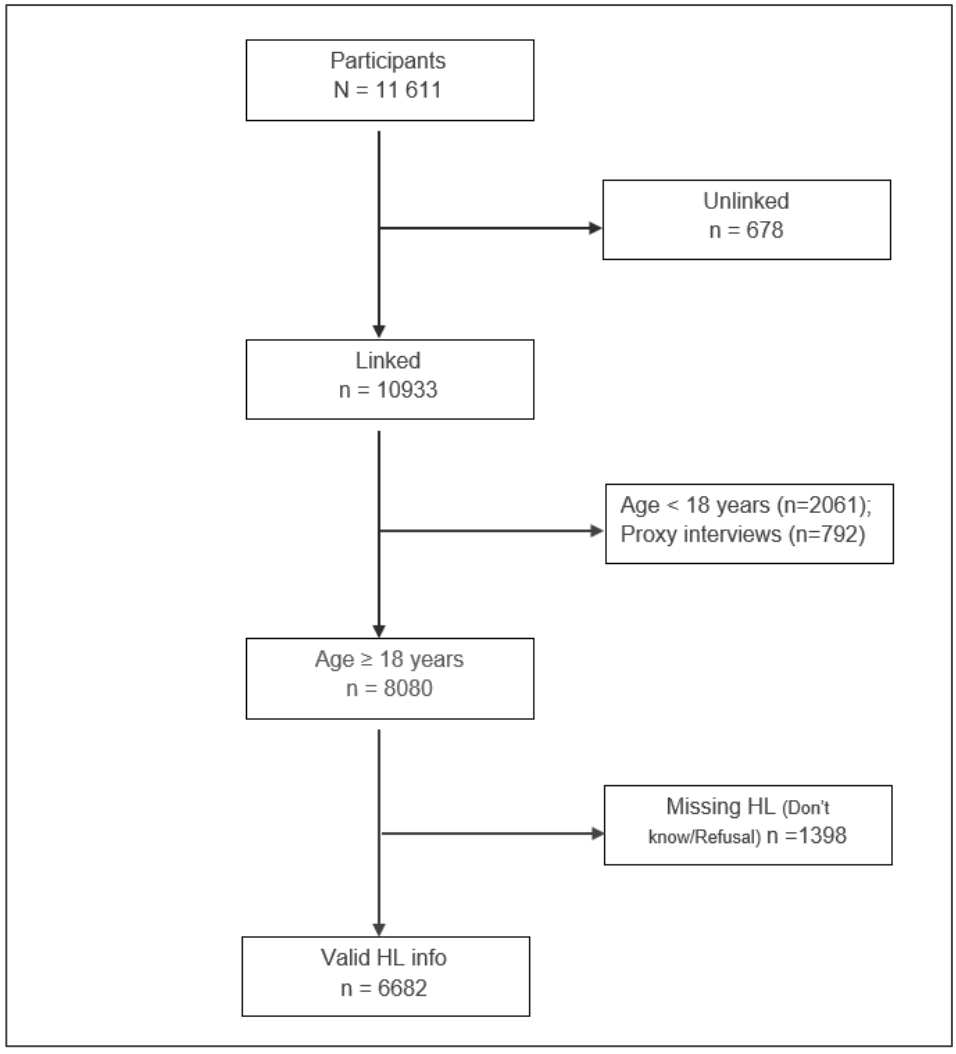


Figure 6.1: Participants' selection process for mediation analysis, HISlink 2018, Belgium

Measures

Dependent variables – Health related outcomes

Health related outcomes included in this study were either retrieved from the BCHI data (preventive dental care use, breast cancer screening, vaccination against flu among older people, purchase of antibiotics and antidepressants) or from the BHIS data (perceived health status, physical activity, diet, alcohol and tobacco consumption). The purchase of antidepressants was used as a proxy for mental health status. A detailed variable description and operationalization is found in Table 6.1.

Independent variables

Information regarding the independent variables was included from the BHIS. Educational attainment and income were utilized as proxy indicators for SES. These variables have frequently been used as indicators of SES in previous studies (8,9,12,13,26). Other indicators such as occupation (9,26) and race/ethnicity (8,13) were not considered here.

Mediator variable

The HL level of the Belgian population was assessed via the Belgian BHIS in 2018, using the 6-items European Health Literacy Survey Questionnaire (HLS-EU-Q6), a short-short form of the original 47-items HL questionnaire (HLS-EU-Q47) (27). Like the original, the HLS-EU-Q6 is a self-reported tool for which participants are asked to indicate how easy or difficult they find it to perform an information-related task (e.g., "judge when you may need to get a second opinion from another doctor", "use information the doctor gives you to make decisions about an illness"), using Likert-type responses. Detailed information on the construction of the HL level is found in Table 6.1. Based on the final score, three possible levels of HL are defined: insufficient, limited and sufficient level of HL. In this study, HL was treated as a dichotomous variable grouping together insufficient and limited as insufficient HL vs. sufficient HL.

Confounding variables

Based on previous studies, the demographic characteristics that were identified as potential confounders in the assessment of the association between SES and health outcomes were sex (male/female) and age (in years as a continuous variable) (9,10,12,26).

Table 6.1: Variables description and operationalization, HISlink 2018, Belgium

Variables name	Variable description / operationalisation
Dependent variables – Health related outcomes	
Preventive dental care among adult population aged 18 years and over	The selected indicator is the proportion of the adult population aged 18 years and over who had at least one contact with a dentist in the reference period, i.e. in 2018, for preventive care such as an oral examination, a prophylactic cleaning, scaling, etc. The specific NIHDI nomenclature codes for the preventive dental care can be found in (28).
Purchase of antibiotic among population aged 18 years and over	This indicator is defined as the proportion of the population aged 18 years and over with at least one purchase of antibiotics between 01/07/2018 and 30/06/2019. Pharmanet data were used to identify cases of purchase of antibiotics. Purchase of a prescribed antibiotic was defined as having obtained at least one reimbursement of prescribed medicine belonging to ATC-code group J01 (antibacterials for systemic use) purchased from a public pharmacy (see Table A1 in the supplementary file). As antibiotic purchase has probably a seasonal pattern, there may be more than one peak in antibiotics use in a calendar year. Therefore in order to include only one winter peak per 12-month period, instead of the months January to December, we used the period from July 01, 2018 to June 30, 2019 to express the annual antibiotic purchase (29).
Vaccination against flu among community dwelling older people aged 65 years and older	The indicator expresses the proportion of the population aged 65 years and over that is vaccinated against flu in the reference period, i.e., calendar year 2018. Older people aged 65 years and over residing in an institution (rest homes and the rest and care homes) were excluded because in the BCHI data only vaccines which have been reimbursed are taken into account and since 2010 vaccines are free of charge for older people residing in an institution in Flanders (30). Hence the calculations for this indicator may result in an underestimation of the true coverage rate. All vaccines belonging to the ATC 4 class J07BB (anti-influenza vaccines) were considered.
Mental health	The purchase of antidepressants is used as a proxy of mental health. The indicator expresses the percentage of adults aged 18 years and over with at least one purchase of an antidepressant (30) (ATC code=N06A) in 2018.
Breast cancer screening among women aged 50-69 year in 2018	Proportion of women aged 50-69 having received at least one mammogram within the last two years, i.e., within the reference year or the reference year-1. In the BCHI data source, the mammographies realized within the screening programme follow a specific procedure, and have their own billing codes. However, these codes do not allow to sufficiently discriminate screening within the program from the other mammographies (opportunistic screening, diagnostic evaluation). Therefore, in this study, all mammograms are considered, within or outside the context of the organised screening programme and we assumed that the largest part of the mammographies undergone between 50 and 69 is made for screening purposes, and therefore we used this information as a proxy of the breast cancer screening. The NIHDI nomenclature codes used can be found in Table A1 in the supplementary file.
Perceived health status *among population aged 18 years and over*	Self-perceived health or self-rated health (SRH) is based on the single question: “How is your health in general?”. This question is part of the Minimum European Health Module (MEHM), which is internationally used. Five response categories are possible: Very good / Good / Fair / Poor / Very poor. The response categories Very good / Good are recorded as “Good” and those Fair / Poor / Very poor as “Poor”.
Physical activity among population	This refers to non-work-related physical activity (leisure-time physical activity and/or the use of a bicycle for commuting) meeting WHO recommendations: spend at least 150 minutes per week in physical activities of at least moderate intensity. The Physical Activity Questionnaire developed by European Health

aged 18 years and over	Interview Survey (EHIS-PAQ) was used to assess physical activity. This is a dichotomous variable (Practice of physical activity / No practice of physical activity).
Type of diet among population aged 18 years and over	The type of diet was assessed using a short food frequency questionnaire. The indicator refers to the proportion of the population aged 18 years and over who eat the recommended daily amount of fruit and vegetables, i.e., at least 5 portions fruits and vegetables (Healthy diet) or not (Unhealthy diet).
Consumption of alcohol among population aged 18 years and over	The EHIS wave 3 questions (31) are used to measure alcohol consumption in order to comply to the European Regulation which recommends the use of a harmonised approach in all EU Member States. The indicator expresses the drinking frequency (at least once a week/less than once a week) among the population aged 18 years.
Consumption of tobacco among population aged 18 years and over	*Proportion of the population aged 18 and over who currently smoke (daily or occasionally).* The tobacco consumption is a dichotomous variable (Yes / No).
Independent variables	
Educational attainment	Educational attainment is based on the highest level of education achieved in the household. Possible values are “primary or no degree”, “secondary inferior”, “secondary superior”, and “superior education” following the ISCED-11 classification, whereby superior education includes all obtained degrees higher than secondary superior (32). These values are recorded into two categories for the analyses: lower secondary’s degree or lower (“primary or no degree”, “secondary inferior”) and higher secondary’s degree or higher (“secondary superior”, and “superior education”).
Household income level	The quintiles of the equivalent household income (quintile 1: <750, quintile 2: 751-1000, quintile 3: 1001-1500, quintile 4: 1501-2500, quintile 5: >2500) were recoded in low (quintile 1–3) and high (quintile 4 and 5).
Mediator variable	
Health literacy (HL) among population aged 18 years and over	The HL level was assessed in the BHIS 2018, using the 6-items European Health Literacy Survey Questionnaire (HLS-EU-Q6), which is a short- form of the original 47-items tool (HLS-EU-Q47) (27). Like the original, the HLS-EU-Q6 is a self-reported tool whereby participants are asked how easy or difficult they find it to perform an information-related task, using Likert-type responses (“very easy” = 4; “fairly easy” = 3; “fairly difficult” = 2; “very difficult” = 1. “Don’t know” or refusal were recoded as missing. The six items covered are: • Judge when you may need to get a second opinion from another doctor • Use information the doctor gives you to make decisions about an illness • Find information on how to manage certain mental health problems like stress or depression • Judge if the information on health risks in the media is reliable? (Examples: TV, Internet or other media) • Find out about activities that are good for your mental well-being? (Examples: meditation, sport, walking,...) • Understand information in the media on how to get healthier? (Examples: Internet, newspapers, magazines). The scale final score measuring HL is the mean value on the six items, which varies between 1 and 4. Only respondents who answered at least 5 items were considered. Based on the final score, three possible levels of HL are defined: insufficient level of HL ($1 \leq x \leq 2$); limited level of HL ($2 < x < 3$); sufficient level of HL ($3 \leq x \leq 4$). In this study, HL was a dichotomous variable

	grouping together insufficient and limited levels of HL as “insufficient HL” - vs. ”sufficient level of HL”.
Confounding variables	
Age	Respondents age (in years)
Sex	Respondents sex (Male / Female)

Statistical analysis

Descriptive analysis

Descriptive statistics summarizing the sociodemographic characteristics of the participants are presented as a percentage in case of categorical variables and as a mean in case of continuous variables. Participants’ characteristics were estimated overall and by level of HL. Comparisons were statistically tested using a $\chi 2$ test for categorical variables and a t-test for normally distributed continuous age variable. Correlation analyses were performed to determine the relationships between the main variables, i.e., the independent variables, the outcomes variables and the mediator variable (10–12,33,34) prior to mediation analysis. Table A2 in the supplementary file provides the guidance of correlation coefficient interpretation (35). In addition, the association between SES and health related outcomes was tested after controlling for age and sex in a regression analysis. Only associations that remained significant after adjusting for age and sex were considered for the mediation analysis.

Mediation analysis

To test the hypothesis that HL is a pathway through which educational attainment and household income affect the selected health related outcomes, the mediation effect of HL was examined separately for each of the two SES factors considered (9) and for each of the selected outcomes.

The analysis proceeded in two steps. First, two logistic regression models were specified: (1) the mediator model for the conditional distribution of the mediator (HL) given the independent variable (SES), and (2) the outcome model for the conditional distribution of the outcomes given the independent variable and the mediator. These models were fitted separately and controlled for age and sex as covariates (except for breast cancer screening where the model was controlled for age only) because they were expected to be all related to the key variables (see Figure 6.2 for the conceptual model). Age was entered as a continuous variable, whereas sex, HL and SES were

dichotomous variables (9). The outcome model also contained an interaction term for the independent variables x the mediator (9,36). By including an interaction term, we assume that the odds ratio (OR) comparing categories of SES differs according to the mediator variable, i.e., HL, and vice versa. The outputs from the mediator and outcome regression models served as the main inputs to estimate the causal effects for the single mediator model (9,36–38).

A sensitivity analysis was carried out for the purchase of antidepressants (using a threshold of 90 DDD per year of specific medication ATC codes to take into account the quantity of antidepressants purchased).

All analyses were performed using SAS® (version 9.4), taking into account the survey weights for the descriptive analysis. The Causalmed procedure was used for the mediation analysis (37,39). Bootstrap methods (1000 bootstrapped samples) were used to compute standard errors and confidence intervals for causal mediation effects and decompositions (10,11,39,40). The Causalmed procedure computes the total effect of the independent variable on the outcome and decomposes this effect into the indirect and direct effects (39). In terms of interpretation, the indirect effect reflects the magnitude of the effect that is transmitted through the mediator, whereas the direct effect accounts for all the other possible causal chains. Furthermore, the Causalmed procedure yields the proportion mediated, which should be interpreted as an estimate of the percentage of the total effect that is exerted through the mediator (12,26,37,39) and provide insight into the relative importance of the mediating role of HL. Missing values were excluded from the analyses. For each analysis, an α level below 0.05 was considered as significant. All P values are two-tailed.

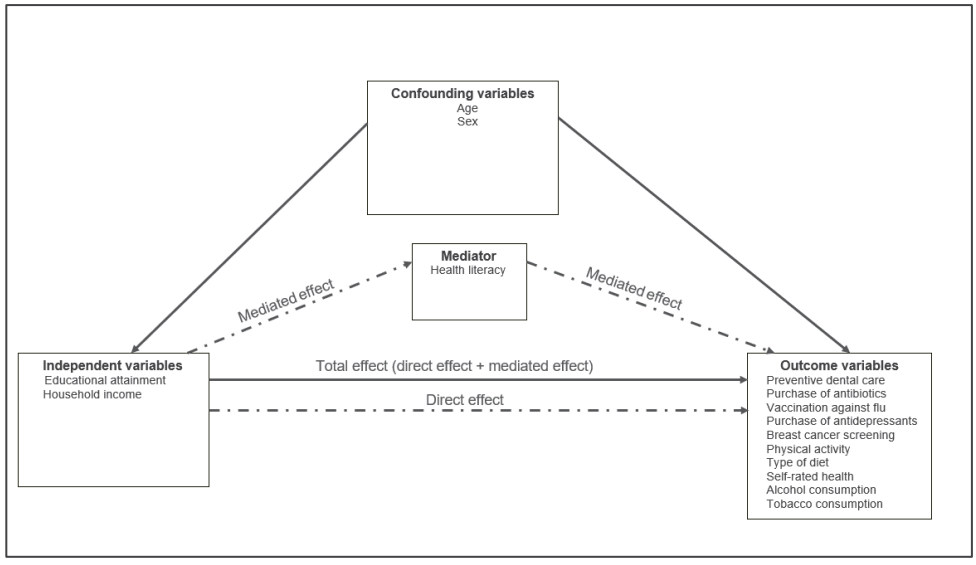


Figure 6.2: Conceptual model of HL as a mediator of the association between SES factors and health related outcomes, HISlink 2018, Belgium.

6.4. RESULTS

Descriptive statistics

Participants characteristics

Participants characteristics are presented in Table 6.2. The crude *n* are presented but all percentages are weighted to match the distribution of the population in terms of age, sex and region of residence. Females represented 51.7% of the adult population and the mean age is 49.4 years old (SD = 0.3). More than eight participants out of ten obtained a higher secondary degree or higher (82.8%). As for income, 48.7% of the participants belonged to a household with higher income category. In terms of HL, sufficient level of HL was found in 66.6% of the population. People who had a sufficient level of HL were more likely to be male, higher educated, and belong to a high income household. Further characteristics are found in Table 6.2.

Table 6.2: Participants characteristics overall and by level of health literacy, n = 6682, HISlink 2018, Belgium

	Total n (% for column)	Health Literacy (HL) levels n (% for raw)		P value
		Sufficient level of HL	Insufficient level of HL	
All	6682 (100)	4411 (66.6)	2271 (33.4)	
Sex				*0.0141*
Male	3161 (48.3)	2116 (68.0)	1045 (32.0)	
Female	3521 (51.7)	2295 (65.2)	1226 (34.8)	
Age, mean ± SE	49.4 ± 0.3	49.4 ± 0.4	49.4 ± 0.6	*0.9414*
Educational attainment				
Lower secondary degree or lower	1144 (16.0)	592 (51.4)	552 (48.6)	<0.0001
Higher secondary degree or higher	5431 (82.8)	3749 (69.5)	1682 (30.5)	
Missing	107 (1.2)	70 (68.7)	37 (31.3)	
Income				<0.0001
Lower income	2769 (39.0)	1712 (61.4)	1057 (38.6)	
Higher income	3012 (48.7)	2097 (70.0)	915 (30.0)	
Missing	901 (12.3)	602 (69.6)	299 (30.4)	

Prevalence of health outcomes

Figure 6.3 illustrates the prevalence of the health related outcomes. The prevalences are described below by domain.

Health behaviours

Among the population studied, 32.2% were physically active, 14.0% reported a healthy diet, 19.8% were current smokers and more than one out of two (52.2%) drank alcohol at least once a week in the 12 months prior to the BHIS data collection. Individuals with insufficient HL were less likely to be physically active, to drink alcohol weekly, and to have a healthy diet, but were more likely to be current smokers than those with sufficient level of HL.

Health status

More than three quarters (76.7%) of the population reported a good perceived health status, while 13.0% had poor mental health. Individuals with insufficient HL were less likely to report good perceived health but more likely to report poor mental health than those with sufficient level of HL.

Use of medicine

In this domain, 36.2% of the population purchased an antibiotic at least once from July 01, 2018 to June 30, 2019. People with insufficient level of HL were more likely to purchase an antibiotic than those with sufficient HL.

Use of preventive health care

Of the participants, 40.0% had a preventive dental care consultation in 2018. As far as flu vaccination is concerned, 55.8% of the population aged 65 years and older were vaccinated in 2018. Finally, 66.1% of women aged 50-69 years were screened for breast cancer in the past two years. Individuals with insufficient level of HL were less likely to have a preventive dental care consultation. In contrast, they were more likely to be vaccinated against flu. There was no significant difference regarding breast cancer screening.

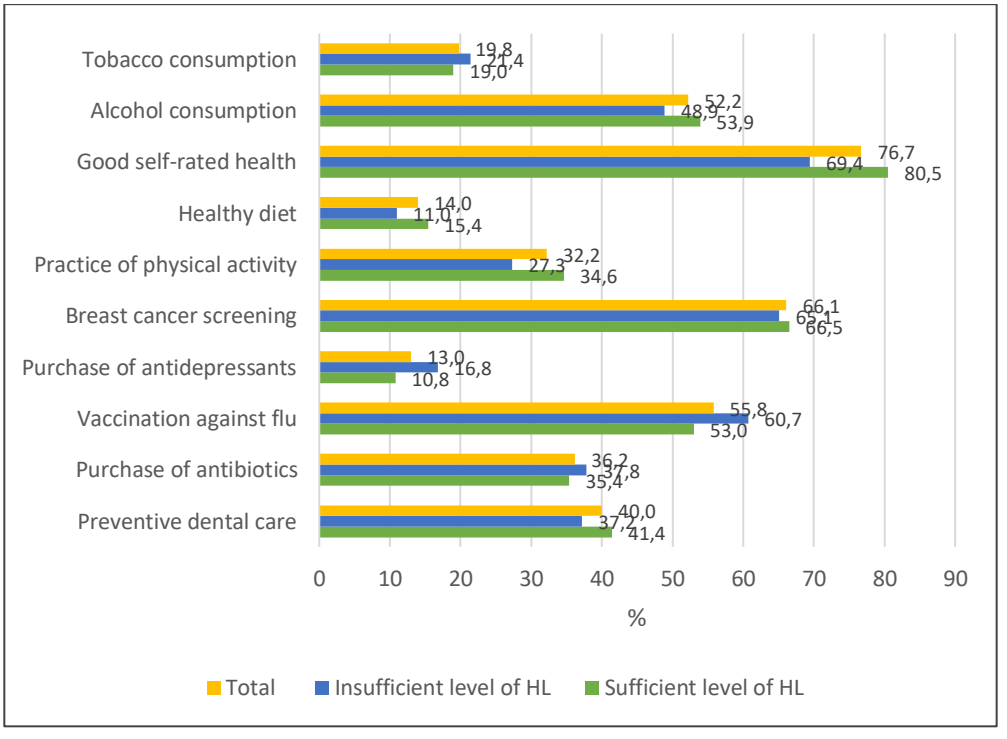


Figure 6.3: Prevalence of health related outcomes overall and by HL level, HISlink 2018, Belgium

Association between health literacy, educational attainment, household income and health related outcomes

The results of the unadjusted association (correlation analysis) are presented in Table A3 (Supplementary file). Only the results from regression analysis adjusted for confounding factors are presented below.

Association between HL and SES

Lower educational attainment and to a lesser extent lower income are associated with having an insufficient level of HL (Table 6.3).

Table 6.3: Association between HL (Insufficient level of HL" vs. "Sufficient level of HL) and independent variables

	Odds Ratio[a] (95% CI)
Educational attainment	
Lower secondary degree or lower	2.19 (1.91-2.50)***
Higher secondary degree or higher	1
Income category	
Lower income	1.45 (1.29-1.62)***
Higher income	1

*[a] Adjusted by age and sex, for breast cancer screening, the OR is adjusted for age only; ** p < 0.05; *** p < 0.0001*

Association between SES and health related outcomes

Association between SES and health behaviour

Lower educational attainment and lower income are associated with lower likelihood of being physically active, having a healthy diet, and reporting weekly alcohol consumption. In contrast, lower educational attainment and lower household income are associated with a higher likelihood of reporting tobacco consumption (Table 6.4 and 6.5 for educational attainment and income respectively).

Association between SES and health status

Lower educational attainment and lower income are associated with a lower likelihood of reporting good perceived health status. Lower income is related to a higher likelihood of having a poor mental health. This association remains significant even

after controlling for education. No significant association is observed between educational attainment and mental health status (Table 6.4 and 6.5 for educational attainment and income respectively).

Association between SES and use of medicine

No significant association is observed between both SES and the purchase of antibiotics (Table 6.4 and 6.5 for educational attainment and income respectively).

Association between SES and use of preventive care

Lower educational attainment and lower income are associated with lower likelihood of receiving preventive dental care and breast cancer screening. No significant association is observed between both SES and vaccination against flu (Table 6.4 and 6.5 for educational attainment and income, respectively).

Association between HL and health related outcomes

HL is positively associated with physical activity, diet and alcohol consumption. In contrast, HL is negatively associated with tobacco consumption. Insufficient level of HL is associated with poor perceived health status and poor mental health status. An insufficient level of HL in the low SES group is associated with a lower likelihood of preventive dental care use. In contrast, insufficient level of HL is associated with a greater likelihood of vaccination against flu. No significant association is observed between HL and purchase of antibiotics and participation in breast cancer screening (Table 6.4 and 6.5 for educational attainment and income model, respectively).

Table 6.4: Association between health literacy, educational attainment and health related outcomes, HISlink 2018, Belgium

	Odds Ratio[a] (95% CI)									
	Health behaviour				**Health status**		**Use of medicine**	**Preventive health care**		
	Physical activity	Healthy diet	Alcohol consumption (At least once a week)	Tobacco consumption (Current smokers)	Good self-rated health	Poor mental health	Purchase of antibiotics	Preventive dental care	Vaccination against flu	Breast cancer screening
Health literacy										
Insufficient level of HL	0.79 (0.69-0.89)**	0.72 (0.61-0.86)**	0.93 (0.82-1.04)	1.28 (1.11-1.48)**	0.55 (0.48-0.63)***	1.62 (1.37-1.91)***	1.08 (0.96-1.22)	1.00 (0.89-1.12)	1.35 (1.01-1.79)**	0.99 (0.74-1.32)
Sufficient level of HL	1	1	1	1	1	1	1	1	1	1
Educational attainment										
Lower secondary degree or lower	0.62 (0.50-0.77)***	0.59 (0.45-0.77)**	0.42 (0.35-0.51)***	1.71 (1.38-2.13)***	0.61 (0.50-0.75)***	1.08 (0.83-1.40)	1.10 (0.92-1.33)	0.47 (0.38-0.57)***	1.12 (0.83-1.52)	0.59 (0.41-0.85)**
Higher secondary degree or higher	1	1	1	1	1	1	1	1	1	1
Educational attainment and HL interaction term	0.67 (0.47-0.95)**	0.69 (0.44-1.08)	0.75 (0.57-0.99)**	0.84 (0.61-1.16)	0.82 (0.62-1.09)	0.91 (0.64-1.30)	1.07 (0.82-1.40)	0.70 (0.51-0.95)**	0.67 (0.42-1.07)	0.76 (0.42-1.37)

[a] *Adjusted by age and sex, for breast cancer screening, the OR is adjusted for age only; ** p < 0.05; *** p < 0.0001*

Table 6.5: Association between health literacy, income and health related outcomes, HISlink 2018, Belgium

Odds Ratio[a] (95% CI)										
	Health behaviour				**Health status**		**Use of medicine**	**Preventive health care**		
	Physical activity	Healthy diet	Alcohol consumption (At least once a week)	Tobacco consumption (Current smokers)	Good self-rated health	Poor mental health	Purchase of antibiotics	Preventive dental care	Vaccination against flu	Breast cancer screening
Health literacy										
Insufficient level of HL	0.76 (0.64-0.90)**	0.67 (0.53-0.84)**	0.84 (0.72-0.99)**	1.24 (1.01-1.52)**	0.54 (0.44-0.67)***	1.50 (1.17-1.92)**	1.09 (0.92-1.28)	0.99 (0.84-1.15)	1.84 (1.17-2.89)**	0.90 (0.59-1.39)
Sufficient level of HL	1	1	1	1	1	1	1	1	1	1
Income category										
Lower income	0.71 (0.61-0.82)***	0.81 (0.67-0.97)**	0.53 (0.47-0.61)***	1.68 (1.42-2.00)***	0.50 (0.43-0.59)***	1.49 (1.22-1.83)**	1.03 (0.90-1.18)	0.62 (0.54-0.71)***	1.06 (0.80-1.41)	0.52 (0.38-0.71)***
Higher income	1	1	1	1	1	1	1	1	1	1
Income and HL interaction term	0.85 (0.66-1.10)	0.93 (0.66-1.30)	1.03 (0.82-1.29)	1.01 (0.77-1.33)	0.93 (0.72-1.21)	1.03 (0.75-1.43)	1.05 (0.84-1.32)	0.80 (0.64-0.99)**	0.52 (0.31-0.89)	1.03 (0.59-1.79)

*[a] Adjusted by age and sex, for breast cancer screening, the OR is adjusted for age only; ** p < 0.05; *** p < 0.0001.*

Mediation effect of health literacy

Mediation effect of HL on the relationship between educational attainment and health related outcomes

Table 6.6 presents the results of mediation analysis.

Health behaviour

On average, HL is found to significantly mediate the associations between educational attainment and all the health behaviours considered except tobacco consumption, i.e., physical activity (OR of indirect effect = 0.90, 95% CI: 0.85-0.95), diet (OR of indirect effect = 0.89, 95% CI: 0.83-0.95), and alcohol consumption (OR of indirect effect = 0.94, 95% CI: 0.90-0.98). The percentage mediated varies between 3.9% for alcohol consumption to 12.1% and 11.7% for physical activity and diet, respectively.

Health status

HL mediates the association between educational attainment and perceived health status (OR of indirect effect = 0.88, 95% CI: 0.84-0.91), accounting for 15.0% of the total effect.

Preventive health care

A significant mediating role of HL is found for the relationship between educational attainment and preventive dental care (OR of indirect effect = 0.94, 95% CI: 0.89-0.99). This effect accounts for 4.4% of the variance.

Table 6.6: Mediation effects of health literacy (reference = sufficient level of health literacy) in the relationship between health related outcomes[a] and educational attainment (reference = higher secondary degree or higher), HISlink 2018, Belgium

	Odds Ratio[b] (95% CI)
Health behaviour	
Practice of physical activity vs. No practice of physical activity	
Total Effect	0.51 (0.42-0.60)***
Direct effect	0.57 (0.47-0.68)***
Indirect effect	0.90 (0.85-0.95)***
Percentage mediated (%)	12.1 (5.4 to 21.4)**
Healthy diet vs. Unhealthy diet	
Total Effect	0.49 (0.38-0.60)***
Direct effect	0.55 (0.43-0.69)***
Indirect effect	0.89 (0.83-0.95)**
Percentage mediated (%)	11.7 (4.2 to 23.1)**
Alcohol consumption (At least once a week vs. Less than once a week)	
Total Effect	0.37 (0.32-0.42)***
Direct effect	0.39 (0.34-0.45)***
Indirect effect	0.94 (0.90-0.98)**
Percentage mediated (%)	3.9 (1.4-7.0)**
Tobacco consumption (current smokers vs. No current smokers)	
Total Effect	1.64 (1.37-1.93)***
Direct effect	1.62 (1.34-1.91)***
Indirect effect	1.01 (0.96-1.07)
Percentage mediated (%)	3.5 (-9.8 to 17.2)
Health status	
Good self-rated health vs. Poor self-rated health	
Total Effect	0.52 (0.44-0.61)***
Direct effect	0.59 (0.50-0.71)***
Indirect effect	0.88 (0.84-0.91)***
Percentage mediated (%)	15.0 (8.5-26.6)**
Preventive health care	
Preventive dental visit vs. No preventive dental visit	
Total Effect	0.40 (0.34-0.47)***
Direct effect	0.43 (0.36-0.50)***
Indirect effect	0.94 (0.89-0.99)**
Percentage mediated (%)	4.4 (0.9 to 8.7)**

a Certain health related outcomes were not included because after controlling for confounding factors, the association between these outcomes and education (mental health status, purchase of antibiotics, vaccination against flu) or health literacy (purchase of antibiotics, breast cancer screening) was no longer

*significant; b Adjusted by age and sex, for breast cancer screening, the OR is adjusted for age only; Bootstrap Percentile 95% Confidence Limits; ** p < 0.05; *** p < 0.0001. All P values are two-tailed.*

Mediation effect of HL in the relationship between income and health related outcomes

Table 6.7 presents the results of mediation analysis.

Health behaviour

HL significantly mediates the association between income and physical activity (OR of indirect effect = 0.97, 95% CI: 0.95-0.98), diet (OR of indirect effect = 0.96, 95% CI: 0.94-0.98) and tobacco consumption (OR of indirect effect = 1.02, 95% CI: 1.01-1.04). The percentage mediated ranges from 4.6% to 12.1%. No mediating role of HL is found for the relationship between income and alcohol consumption.

Health status

A mediating role of HL is found for the association between income and perceived health status (OR of indirect effect = 0.95, 95% CI: 0.93-0.97). The indirect effect accounts for 4.5% of the total effect. HL significantly mediates the association between income and mental health status (OR of indirect effect = 1.04, 95% CI: 1.02-1.07)., accounting for 10.4% of the total effect of income. In sensitivity analysis, even taking into account a threshold of 90 DDD of antidepressants, the mediating effect of HL in the relationship between income and mental health status remains significant (OR of indirect effect = 1.04, 95% CI: 1.02-1.07). The percentage mediated is about 12.7% (see Table A4 in the supplementary file).

Preventive health care use

HL acts as mediator in the relationship between income and use of preventive dental care, (OR of indirect effect = 0.98, 95% CI: 0.97-0.99), accounting for 2.5% of the variance.

Table 6.7: Mediation effects of health literacy (reference = sufficient level of health literacy) in the relationship between health related outcomes[a] and household income (reference = higher household income), HISlink 2018, Belgium

	Odds Ratio[b] (95% CI)
Health behaviour	
Practice of physical activity vs. No practice of physical activity	
Total Effect	0.66 (0.58-0.74)***
Direct effect	0.68 (0.60-0.77)***
Indirect effect	0.97 (0.95-0.98)***
Percentage mediated (%)	6.6 (3.2 to 12.0)**
Healthy diet vs. Unhealthy diet	
Total Effect	0.77 (0.65-0.89)**
Direct effect	0.79 (0.67-0.92)**
Indirect effect	0.96 (0.94-0.98)**
Percentage mediated (%)	12.1 (5.0 to 33.0)**
Alcohol consumption (At least once a week vs. Less than once a week)	
Total Effect	0.53 (0.48-0.59)***
Direct effect	0.54 (0.48-0.60)***
Indirect effect	0.99 (0.98-1.00)
Percentage mediated (%)	1.3 (-0.2 to 2.9)
Tobacco consumption (current smokers vs. No current smokers)	
Total Effect	1.73 (1.50-1.99)***
Direct effect	1.69 (1.46-1.95)***
Indirect effect	1.02 (1.01-1.04)**
Percentage mediated (%)	4.6 (1.0 to 9.3)**
Health status	
Good self-rated health vs. Poor self-rated health	
Total Effect	0.47 (0.41-0.55)***
Direct effect	0.50 (0.43-0.57)***
Indirect effect	0.95 (0.93-0.97)***
Percentage mediated (%)	4.5 (2.7 to 7.2)***
Poor mental health status	
Total Effect	1.57 (1.34-1.84)***
Direct effect	1.51 (1.29-1.78)***
Indirect effect	1.04 (1.02-1.07)**
Percentage mediated (%)	10.4 (4.9 to 18.7)**

Preventive health care	
Preventive dental visit vs. No preventive dental visit	
Total Effect	0.57 (0.51-0.64)***
Direct effect	0.58 (0.52-0.65)***
Indirect effect	0.98 (0.97-0.99)**
Percentage mediated (%)	2.5 (0.7 to 4.9)**

*[a] Certain health related outcomes were not included because after controlling for confounding factors, the association between these outcomes and income (purchase of antibiotics, vaccination against flu) or health literacy (purchase of antibiotics, breast cancer screening) was no longer significant ; [b] Adjusted by age and gender, for breast cancer screening, the OR is adjusted for age only; Bootstrap Percentile 95% Confidence Limits; ** p < 0.05; *** p < 0.0001. All P values are two-tailed.*

6.5. DISCUSSION

Main findings

The reduction of SE health disparities is an important objective for public health policies. It is therefore relevant to identify factors that contribute to these disparities. In that regard, HL is of interest as it constitutes a potential pathway through which SES may affect health. Moreover, contrary to structural SES factors that are difficult to modify, HL can be more easily improved (13,26). Indeed, HL can be modified via health and literacy programs while the structural SES factors requires more structural interventions that are beyond the health sector. This study explored whether HL acts as a mediator in the association between SES as measured by educational attainment and household income and the selected health related outcomes that are of interest from a public health perspective.

The SE disparities in health related outcomes are confirmed with our data. HL was found to partly mediate the association between educational attainment and health behaviour (except tobacco consumption) and the association between educational attainment, perceived health status and preventive dental care. HL constitutes a pathway through which income influences health behaviour (except alcohol consumption), perceived health status, mental health status and preventive dental care.

As expected, a mediation effect of HL for the link with SES was found in three out of four of the health behaviours considered. Although the contributing effect of HL to the total effect is rather small, it is in line with the existing evidence (11,21). Indeed, in a

Danish population-based study, Friis et al. (2016) found that HL mediated the relationship between educational attainment and health behavior, especially in relation to being physically inactive (accounting for 5.4% to 20% of the variance depending of the scales from HL questionnaires), having a poor diet (accounting for 13% of the variance), and daily smoking (accounting for 4.5% to 6.6%) (21). Although using different independent variable, Chen et al. (2019) demonstrated that HL played a partial mediating role between social capital and physical activity (8.2% to 12.7% of the total effect) as well as type of diet (4.93% to 12.7% of the total effect) (11).

Compared with the other health behaviours studied, the mediating role of HL in the relationship between SES and alcohol and tobacco consumption is inconsistent. While HL does not appear to mediate the relationship between education and tobacco consumption, as is the case for alcohol consumption, it was found to mediate the relationship between income and tobacco consumption, but not alcohol consumption. Friis et al. (2016) also did not find a mediation effect of HL in the association between education and tobacco consumption. The authors argued that the underlying explanations for this may be link to the fact that in Denmark policy regulations and mass media campaigns relating to tobacco use have been in place for more than two decades. So, regardless of their HL levels, most people are aware of the health-related consequences of smoking (21). A similar result was found by Van Den Broucke et al. (2014) (41). The underlying hypothesis put forward by Friis et al. (2016) could be applied to our findings, because an anti-smoking plan introduced legislative measures in Belgium since 2006 that include, for example, increase in tobacco price, banning smoking in public place and dissuasive colour photos. These measures are likely to have an impact on the risk of individuals' tobacco consumption, whatever their level of HL (42).

The strongest mediation effect of HL was found for the association between educational attainment and perceived health status, suggesting that low educated people manage their health problems less well, resulting in poorer perceived health status. Therefore, a better HL among low educated people will lead to a better perceived health status for them. This result is in line with results from previous studies (8,9,12,13). Some studies have shown that the relative importance of HL as a pathway between education and perceived health status is greater among people with lower levels of education than among those with higher levels of education (9,12), but Van

Heide et al. (2013) also found that the mediating role of HL does not show a linear gradient as education level increases (12). This means that HL exhibited a more important pathway for lower secondary educated than for preprimary/primary educated (12). In the present study, we were unable to explore this issue as we only used two levels of education. To determine the extent to which improving HL could help reduce education-related disparities in health status, further research is needed on the relative importance of the mediating role of health literacy between different levels of education.

As regards mental health status, the association with income is mediated by HL. These results could be explained by the fact that, unlike people with a sufficient level of HL, people with an insufficient level of HL do not know or understand that they can consult a psychologist for their mental health problems and therefore turn to the use of antidepressants. Furthermore, it is less expensive to take antidepressants (which are fully reimbursed) than to undergo therapy (which is not reimbursed).

Finally, with regard to preventive health care, HL significantly mediated the association between both SES and preventive dental care. The vaccination against flu and participation in breast cancer screening were not considered for mediation analysis because after controlling for participants' age and sex, the association between these indicators, SES and/or HL was no longer significant. These findings may be linked to the universal health care system that is in place in Belgium. As suggested by previous studies (26,43), in countries with universal, publicly-funded health care systems, the burden exerted by SES or HL is small or absent, since it is reduced by an equitable access, free of charge, for all the target categories of the population. Therefore, individual decisions are not likely to play a crucial role in this behaviour, and so the influence of HL may be minimal.

Strengths and Limitations

To our knowledge, this is the first study based on the linkage of two population databases to examine whether HL plays a mediating role in the associations between education, income and a number of objective and subjective measures of health-related outcomes in different domains, namely health behaviour, perceived health status, and use of medicine and preventive care in a large sample. Studies most often rely on subjective measures to this respect. However, it has been recognised that to

better understand the association between HL and health outcomes, objective measures of the latter may provide important evidence (12) and should therefore be used wherever possible.

Our study has a number of limitations that must be acknowledged. First, using the criterion of purchasing at least one prescription of antidepressants in the reference period to identify cases of mental health may have caused the inclusion of individuals who use antidepressants for another indication than depression, who did not comply with or respond to the treatment. However, the results from the sensitivity analysis taking into account a threshold of 90 DDD per year of specific medication ATC codes confirmed the mediation effect of HL, meaning that our indicator was accurate. Furthermore, the prevalence of mental health status found in our study is consistent with that found by Van Heide et al. (2013) using self-reported mental health status (12).

A second limitation is that regarding breast cancer screening, no distinction was made between mammographies as part of a screening program and opportunistic mammographies. Even though the mammographies realized within the program have their own billing codes in the BCHI data, they do not sufficiently discriminate screening within the program from the other mammographies (opportunistic screening, diagnostic evaluation). In fact, opportunistic screening mammograms are often miscoded as diagnostic mammograms for reimbursement purposes in the BCHI. However, we assumed that the largest part of the mammographies undergone between 50 and 69 years of age is made for screening purposes, giving information as to preventive health care initiatives.

Third, although the number of missing HL values (17% non-response/refusal) is almost comparable to that reported (13%) in another population-based study (21), this may have affected our results. It is plausible that missingness was higher among people with a low HL level than among people with a high HL level. Therefore, the use of complete case analysis may have affected the final results. Further exploration showed that the missing values are evenly distributed across the HL domains, suggesting an absence of selection in the responses to the six items that make up the HL scale. However, the missing values are not randomly distributed across population subgroups. An exploratory analysis indicated that older people, people with a low level of education, people living in low-income households, those born outside Belgium and

those living in the Brussels-Capital Region are more likely to have missing values on the HL scale (see Table A5 in the supplementary file). Future studies should assess the impact of these missing values as part of a sensitivity analysis using multiple imputation.

Fourthly, the instrument that was used to assess HL in this study was a generic one, which may explain the relatively low percentage of mediated effects that were found. In fact, some authors suggest the use of outcome-specific health literacy instruments (e.g., vaccine literacy) to better assess the role of for decision making in that field (26). However, our instrument is validated and has good validity. The next survey BHIS 2023 includes a more extensive HL instrument (12-item questionnaire) (44,45) and will allow us to verify our findings.

Finally, the dichotomisation of the HL level may has resulted in a loss of information. Dichotomisation puts people with different HL levels in one category and "within differences" in each of the categories are not taken into account in the analysis. This will kind of dilute the information of the HL indicator, as a result of which the mediation effect will be underestimated. The results of this study should therefore be interpreted with caution.

Implications and future perspective

This study has important implications for practitioners and policy makers. Besides the fact that it adds further insights that help to understand the underlying mechanisms linking SES to health related outcomes, the mediating role of HL may have important implications for interventions that are aimed at reducing health disparities, as HL can be modified via health and literacy programs contrary to SES factors. Policies and interventions aimed at increasing the level of HL in the population or that take people's insufficient level HL better into account might effectively contribute to reduce health disparities. As this study again demonstrates, the most vulnerable and disadvantaged people in society are more at risk of limited HL and are known to have the poorest health outcomes. Strategies to improve HL are therefore important empowerment tools which have the potential to reduce health disparities.

Several strategies have been proposed for effective improvement of HL such as:

- developing initiatives to increase the level of HL in the population for example through interventions at several levels (political, institutional, professional,

citizen) (46,47). For example health literacy interventions in the delivery of Medicine in US in which pharmacists, as healthcare professionals who will dispense prescriptions for medication, have a key role to advise the patient on any queries relating to their medication and to counsel on appropriate use. The mental health literacy interventions in adults in which it is assumed that changing mental health literacy will lead to a change in behaviours that benefit mental health, which will, in turn, produce an improvement in mental health (48);

- improving the detection of people with a low level of HL and adapting communication during contact with healthcare professionals;
- creating health literate organisations that incorporate the management of HL into their policies and operations (46);
- covering the entire health continuum and not just the medical aspects;
- following participatory processes aimed particularly at people with a lower level of health literacy (47).

Van den Broucke et al. (2018) highlighted the need to invest in building the capacity of the public health system and of other stakeholders to address health literacy (49). In a similar vein, Public Health England proposed the following strategies (50):

- the use by health and social care services of the simple and effective teach-back method to check user understanding;
- an early intervention approach to health literacy - ensuring that the promotion of health literacy is fully integrated into school and early years curricula, as well as the training of health and social care professionals improving health literacy to reduce health disparities;
- community-based peer support approaches to health literacy that help to spread health literacy through social networks;
- empowerment of professionals through training, continuing education and interdisciplinary initiatives to improve health literacy and strengthen communication between the public and professionals.

It should be noted that in our study, as in other similar studies, in general the influence of HL in the relationship between SES and health related outcomes is rather weak. This may suggest the influence of other factors or mechanisms that need to be investigated. Future research should therefore also take other potential mediators into account, such as social support and environmental exposure. Furthermore, it would be useful to look at mediation effects per stratum (age, sex, cultural background), to allow targeting interventions to specific groups. Zanobini et al. (2022) also suggest to investigate the hypothesis that SES could be the mediator variable between HL and influenza vaccine uptake (26). Finally, since different HL dimensions show distinct direct and indirect pathways in influencing health outcomes (21,51), it is necessary to assess the mediating role of HL separately different dimensions. Based on the findings from such investigation, interventions could be targeting dimensions and population subgroups that are at risk. A multiple mediator models could also be considered (52) for identifying these complex underlying mechanisms.

6.6. CONCLUSIONS

This study provides evidence that HL partially serves as a pathway thorough which educational attainment and income affect health behaviour, perceived health status, mental health status and preventive dental care. Although the mediating influence of HL in this respect is rather limited, the results suggest that strategies to reduce health disparities in these areas could benefit from taking individuals' HL into account in awareness campaigns as part of prevention, patient education and other public health interventions. Further data and analysis are needed to confirm our results and to better explore the mediating effects of HL.

Declarations

Ethics statement

This study was carried out through an individual linkage between the BHIS 2018 data and the BCHI data. The BHIS 2018 was carried out in line with the Belgian privacy legislation and has been approved by the ethics committee of the University hospital of Ghent on December, 21 2017 (advice EC UZG 2017/1454). The participation to the BHIS is voluntary. There was no formal written and signed consent foreseen. The selected households were notified about the survey, its practical organization, the

institution in charge, the commissioners of the survey and its content via a letter and an information leaflet personally addressed to them. It was also clearly stipulated in the letter and the leaflet that participation is voluntary. Participation was equivalent to giving consent. The aforementioned ethics committee of the University hospital of Ghent waived the need for formal written and signed informed consent. Data linkage was authorized by the Information Security Committee (local reference: Deliberation No. 20/204 of November 3, 2020). All methods were performed in accordance with the Declarations of Helsinki.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analysed during the current study are not publicly available because they contain sensitive and identifying information, but are available from the corresponding author on reasonable request. Further information regarding the survey and the data access procedure can be found here: Health Interview Survey | Microdata request procedure | sciensano.be .

Competing interests

The authors declare that they have no competing interests.

Funding

This work did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The Belgian Health Interview Survey (BHIS) is financed by the Federal and Inter-Federated Belgian Public Health authorities. The linkage between BHIS data and the Belgian Compulsory Health Insurance data is financed by the National Institute for Health and Disability Insurance.

Authors' contributions

FB and JVdH were responsible for designing the objectives and approach of the study. FB conducted the literature searches and summaries of previous related work, undertook the statistical analyses in collaboration with JVdH, interpreted the results, wrote the initial version of the manuscript and conducted the revisions. FB was the

main contributor in writing the manuscript. JVdH, LG, SD, RC, SVdB, and OB critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Statbel, the Belgian statistical office, which was responsible for BHIS sample selection and fieldwork management. Thanks to Statbel and the InterMutualistic Agency (IMA) for their involvement in the process of data linkage. And of course, all the people who voluntarily participated in the Belgian health interview survey.

6.7. BIBLIOGRAPHY

1. Kulik MC, Menvielle G, Eikemo TA, Bopp M, Jasilionis D, Kulhánová I, et al. Educational Inequalities in Three Smoking-Related Causes of Death in 18 European Populations. Nicotine & Tobacco Research. 2014 May 1;16(5):507–18.
2. Adler NE, Newman K. Socioeconomic Disparities In Health: Pathways And Policies. Health Affairs. 2002 Mar;21(2):60–76.
3. Marmot M, Allen J, Bell R, Bloomer E, Goldblatt P. WHO European review of social determinants of health and the health divide. The Lancet. 2012 Sep;380(9846):1011–29.
4. Kröger H, Pakpahan E, Hoffmann R. What causes health inequality? A systematic review on the relative importance of social causation and health selection. Eur J Public Health. 2015 Dec;25(6):951–60.
5. Paasche-Orlow MK, Wolf MS. The Causal Pathways Linking Health Literacy to Health Outcomes. am j health behav. 2007 Jul 1;31(1):19–26.
6. Husson O, Mols F, Fransen MP, van de Poll-Franse LV, Ezendam NPM. Low subjective health literacy is associated with adverse health behaviors and worse health-related quality of life among colorectal cancer survivors: results from the profiles registry: Health literacy and health outcomes. Psycho-Oncology. 2015 Apr;24(4):478–86.
7. Schillinger D, Barton LR, Karter AJ, Wang F, Adler N. Does Literacy Mediate the Relationship between Education and Health Outcomes? A Study of a Low-Income Population with Diabetes. Public Health Rep. 2006 May;121(3):245–54.
8. Mantwill S, Monestel-Umaña S, Schulz PJ. The Relationship between Health Literacy and Health Disparities: A Systematic Review. Antonietti A, editor. PLoS ONE. 2015 Dec 23;10(12):e0145455.
9. Lastrucci V, Lorini C, Caini S, Florence Health Literacy Research Group, Bonaccorsi G. Health literacy as a mediator of the relationship between socioeconomic status and health: A cross-sectional study in a population-based sample in Florence. Kwon YD, editor. PLoS ONE. 2019 Dec 23;14(12):e0227007.
10. Shih YL, Hsieh CJ, Lin YT, Wang YZ, Liu CY. The Mediation Effect of Health Literacy on Social Support with Exchange and Depression in Community-Dwelling Middle-Aged and Older People in Taiwan. Healthcare. 2021 Dec 19;9(12):1757.
11. Chen WL, Zhang CG, Cui ZY, Wang JY, Zhao J, Wang JW, et al. The impact of social capital on physical activity and nutrition in China: the mediating effect of health literacy. BMC Public Health. 2019 Dec;19(1):1713.
12. van der Heide I, Wang J, Droomers M, Spreeuwenberg P, Rademakers J, Uiters E. The Relationship Between Health, Education, and Health Literacy: Results

From the Dutch Adult Literacy and Life Skills Survey. Journal of Health Communication. 2013 Dec 4;18(sup1):172–84.

13. Stormacq C, Van den Broucke S, Wosinski J. Does health literacy mediate the relationship between socioeconomic status and health disparities? Integrative review. Health promotion international. 2019;34(5):e1–17.
14. Sørensen K, Pelikan JM, Röthlin F, Ganahl K, Slonska Z, Doyle G, et al. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). Eur J Public Health. 2015 Dec;25(6):1053–8.
15. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low Health Literacy and Health Outcomes: An Updated Systematic Review. Ann Intern Med. 2011 Jul 19;155(2):97.
16. Vandenbosch J, Van den Broucke S, Vancorenland S, Avalosse H, Verniest R, Callens M. Health literacy and the use of healthcare services in Belgium. J Epidemiol Community Health. 2016 Oct;70(10):1032–8.
17. Bostock S, Steptoe A. Association between low functional health literacy and mortality in older adults: longitudinal cohort study. BMJ. 2012 Mar 15;344(mar15 3):e1602–e1602.
18. Kickbusch I, Pelikan JM, Apfel F, Tsouros AD, World Health Organization, editors. Health literacy: the solid facts. Copenhagen: World Health Organization Regional Office for Europe; 2013. 73 p. (The solid facts).
19. On behalf of the Preventive Evidence into Practice (PEP) Partnership Group, Jayasinghe UW, Harris MF, Parker SM, Litt J, van Driel M, et al. The impact of health literacy and life style risk factors on health-related quality of life of Australian patients. Health Qual Life Outcomes. 2016 Dec;14(1):68.
20. Geboers B, Reijneveld SA, Jansen CJM, de Winter AF. Health Literacy Is Associated With Health Behaviors and Social Factors Among Older Adults: Results from the LifeLines Cohort Study. Journal of Health Communication. 2016 Aug;21(sup2):45–53.
21. Friis K, Lasgaard M, Rowlands G, Osborne RH, Maindal HT. Health literacy mediates the relationship between educational attainment and health behavior: a Danish population-based study. Journal of Health Communication. 2016;21(sup2):54–60.
22. Rondia K, Adriaenssens J, Van Den Broucke S, Kohn L. Health literacy: what lessons can be learned from the experiences of other countries? [Internet]. Brussels, Belgium: KCE; 2019 [cited 2023 Oct 17]. Report No.: KCE Report 322. Available from: https://kce.fgov.be/sites/default/files/2021-11/KCE_322_Health_Literacy_Report.pdf
23. Svendsen MT, Bak CK, Sørensen K, Pelikan J, Riddersholm SJ, Skals RK, et al. Associations of health literacy with socioeconomic position, health risk behavior, and health status: a large national population-based survey among Danish adults. BMC public health. 2020;20(1):1–12.

24. Demarest S, Van der Heyden J, Charafeddine R, Drieskens S, Gisle L, Tafforeau J. Methodological basics and evolution of the Belgian health interview survey 1997–2008. Arch Public Health. 2013 Dec;71(1):24.

25. Agence InterMutualiste -InterMutualistisch Agentschap (AIM-IMA). Agence InterMutualiste -InterMutualistisch Agentschap [Internet]. [cited 2021 Jul 26]. Available from: https://www.ima-aim.be/-Donnees-de-sante-

26. Zanobini P, Lorini C, Caini S, Lastrucci V, Masocco M, Minardi V, et al. Health Literacy, Socioeconomic Status and Vaccination Uptake: A Study on Influenza Vaccination in a Population-Based Sample. IJERPH. 2022 Jun 6;19(11):6925.

27. Pelikan JM, Röthlin F, Ganahl K. Measuring comprehensive health literacy in general populations: validation of instrument, indices and scales of the HLS-EU study. In Bethesda, Maryland; 2014 [cited 2022 Jun 1]. Available from: https://www.bumc.bu.edu/healthliteracyconference/files/2014/06/Pelikan-et-al-HARC-2014-fin.pdf

28. NIHDI. For a healthy Belgium Medical Practive Variations Preventive dental care - Data [Internet]. [cited 2023 Jan 24]. Available from: https://www.healthybelgium.be/en/medical-practice-variations/digestive-system/dentistry/preventive-dental-care

29. Coenen S, Gielen B, Blommaert A, Beutels P, Hens N, Goossens H. Appropriate international measures for outpatient antibiotic prescribing and consumption: recommendations from a national data comparison of different measures. Journal of Antimicrobial Chemotherapy. 2014 Feb 1;69(2):529–34.

30. Devos C, Cordon A, Lefèvre M, Obyn C, Renard F, Bouckaert N, et al. Performance of the Belgian health system – report 2019 – Supplement. Health Services Research (HSR) [Internet]. Brussels, Belgium: Belgian Health Care Knowledge Centre (KCE); 2020. Report No.: KCE Reports 313S. D/2020/10.273/36. Available from: file:///W:/HIS/HISLINK/HISLINK%202018/References/Mediation%20analysis/Carl%20Devos_2019_2016-06-HSR_Peformance_appendix_technical_2ndedition.pdf

31. Griswold MG, Fullman N, Hawley C, Arian N, Zimsen SRM, Tymeson HD, et al. Alcohol use and burden for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet. 2018 Sep;392(10152):1015–35.

32. UNESCO Institute for Statistics. International standard classification of education: ISCED 2011. Comparative Social Research [Internet]. 2012;30. Available from: http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf

33. Sagong H, Yoon JY. Pathways among Frailty, Health Literacy, Acculturation, and Social Support of Middle-Aged and Older Korean Immigrants in the USA. IJERPH. 2021 Jan 30;18(3):1245.

34. Hai-YanYu, Wu WL, Yu LW, Wu L. Health literacy and health outcomes in China's floating population: mediating effects of health service. BMC Public Health. 2021 Dec;21(1):691.

35. Schober P, Boer C, Schwarte LA. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia. 2018 May;126(5):1763–8.

36. Valeri L. Statistical Methods for Causal Mediation Analysis. A dissertation presented to The Department of Biostatistics in partial fulfilment of the requirements for the degree of Doctor of Philosophy in the subject of Biostatistics. [Internet]. Harvard University; 2012. Available from: https://dash.harvard.edu/bitstream/handle/1/10403677/Valeri_gsas.harvard_0084L_10690.pdf?sequence=3

37. Valente MJ, Rijnhart JJM, Smyth HL, Muniz FB, MacKinnon DP. Causal Mediation Programs in R, M plus, SAS, SPSS, and Stata. Structural Equation Modeling: A Multidisciplinary Journal. 2020 Nov 1;27(6):975–84.

38. VanderWeele TJ, Vansteelandt S. Odds Ratios for Mediation Analysis for a Dichotomous Outcome. American Journal of Epidemiology. 2010 Dec 15;172(12):1339–48.

39. SAS Institute Inc. SAS/STAT®14.3 User's Guide The CAUSALMED Procedur [Internet]. Cary, NC: SAS Institute Inc.; 2017 [cited 2022 Dec 7]. Available from: https://support.sas.com/documentation/onlinedoc/stat/143/causalmed.pdf

40. Wang J, Geng L. Effects of Socioeconomic Status on Physical and Psychological Health: Lifestyle as a Mediator. IJERPH. 2019 Jan 20;16(2):281.

41. Van Den Broucke S, Renwart A. Health literacy mediates the relationship between education level and health behaviour: Stephan Van Den Broucke. European Journal of Public Health. 2014;24(suppl_2):cku164-012.

42. Cellule Générale de Politique Drogues. Stratégie interfédérale 2022-2028 pour une génération sans tabac [Internet]. 2022. Available from: https://organesdeconcertation.sante.belgique.be/sites/default/files/documents/20220323_strategie_interfederale_tabac_note_de_base_fr.pdf

43. Lucyk K, Simmonds KA, Lorenzetti DL, Drews SJ, Svenson LW, Russell ML. The association between influenza vaccination and socioeconomic status in high income countries varies by the measure used: a systematic review. BMC Med Res Methodol. 2019 Dec;19(1):153.

44. Pelikan JM, Link T, Straßmayr C, Waldherr K, Alfers T, Bøggild H, et al. Measuring Comprehensive, General Health Literacy in the General Adult Population: The Development and Validation of the HLS19-Q12 Instrument in Seventeen Countries. IJERPH. 2022 Oct 29;19(21):14129.

45. Duong TV, Aringazina A, Kayupova G, Nurjanah, Pham TV, Pham KM, et al. Development and Validation of a New Short-Form Health Literacy Instrument (HLS-SF12) for the General Public in Six Asian Countries. HLRP: Health Literacy

Research and Practice [Internet]. 2019 Apr [cited 2023 Sep 5];3(2). Available from: https://journals.healio.com/doi/10.3928/24748307-20190225-01

46. Kickbusch I, Pelikan J, Apfel F, Tsouros A. Health literacy: The solid facts. 2013. World Health Organization, Regional Office for Europe. 2013;
47. Brumagne A, Mannaerts D. Littératie en santé: pour une approche globale et émancipatrice visant l'équité. Repères DoRiF. 2018;16.
48. Okan O, Bauer U, Levin-Zamir D, Pinheiro P, Sorensen K. International Handbook of Health Literacy. Policy Press. 2019;
49. Van den Broucke S. Capacity building for health literacy. International Handbook of Health Literacy. 2019;705.
50. Public Health England. Local action on health inequalities: improving health literacy to reduce health inequalities. 2015;
51. Zhang F, Or PP, Chung JW. How different health literacy dimensions influences health and well-being among men and women: The mediating role of health behaviours. Health Expectations. 2021;24(2):617–27.
52. Wang W, Nelson S, Albert JM. Estimation of causal mediation effects for a dichotomous outcome in multiple-mediator models using the mediation formula. Statist Med. 2013 Oct 30;32(24):4211–28.

CHAPTER 7. SUMMARY PAPER

LINKING HEALTH SURVEY DATA WITH HEALTH INSURANCE DATA: METHODOLOGY, CHALLENGES, OPPORTUNITIES AND RECOMMENDATIONS FOR PUBLIC HEALTH RESEARCH. AN EXPERIENCE FROM THE HISLINK PROJECT IN BELGIUM

The findings of this chapter were published as:

Berete F, Demarest S, Charafeddine R, De Ridder K, Van Oyen H, Van Hoof W, Bruyère O, Van der Heyden J. Linking health survey data with health insurance data: methodology, challenges, opportunities and recommendations for public health research. An experience from the HISLink project in Belgium. Archives of Public Health, 2023, 81.1: 198.

Berete *et al. Archives of Public Health* (2023) 81:198
https://doi.org/10.1186/s13690-023-01213-0

Archives of Public Health

COMMENT Open Access

Linking health survey data with health insurance data: methodology, challenges, opportunities and recommendations for public health research. An experience from the HISlink project in Belgium

Finaba Berete[1,2*], Stefaan Demarest[1], Rana Charafeddine[1], Karin De Ridder[1], Herman Van Oyen[1,3], Wannes Van Hoof[1], Olivier Bruyère[4] and Johan Van der Heyden[1]

Abstract

In recent years, the linkage of survey data to health administrative data has increased. This offers new opportunities for research into the use of health services and public health. Building on the HISlink use case, the linkage of Belgian Health Interview Survey (BHIS) data and Belgian Compulsory Health Insurance (BCHI) data, this paper provides an overview of the practical implementation of linking data, the outcomes in terms of a linked dataset and of the studies conducted as well as the lessons learned and recommendations for future links.

Individual BHIS 2013 and 2018 data was linked to BCHI data using the national register number. The overall linkage rate was 92.3% and 94.2% for HISlink 2013 and HISlink 2018, respectively. Linked BHIS-BCHI data were used in validation studies (e.g. self-reported breast cancer screening; chronic diseases, polypharmacy), in policy-driven research (e.g., mediation effect of health literacy in the relationship between socioeconomic status and health related outcomes, and in longitudinal study (e.g. identifying predictors of nursing home admission among older BHIS participants). The linkage of both data sources combines their strengths but does not overcome all weaknesses.

The availability of a national register number was an asset for HISlink. Policy-makers and researchers must take initiatives to find a better balance between the right to privacy of respondents and society's right to evidence-based information to improve health. Researchers should be aware that the procedures necessary to implement a link may have an impact on the timeliness of their research. Although some aspects of HISlink are specific to the Belgian context, we believe that some lessons learned are useful in an international context, especially for other European Union member states that collect similar data.

Keywords Record linkage; data linkage, Health administrative insurance data, Health claims data, Health interview surveys

7.1. ABSTRACT

In recent years, the linkage of survey data to health administrative data has increased. This offers new opportunities for research into the use of health services and public health. Building on the HISlink use case, the linkage of Belgian Health Interview Survey (BHIS) data and Belgian Compulsory Health Insurance (BCHI) data, this paper provides an overview of the practical implementation of linking data, the outcomes in terms of a linked dataset and of the studies conducted as well as the lessons learned and recommendations for future links.

Individual BHIS 2013 and 2018 data was linked to BCHI data using the national register number. The overall linkage rate was 92.3% and 94.2% for HISlink 2013 and HISlink 2018, respectively. Linked BHIS-BCHI data were used in validation studies (e.g. self-reported breast cancer screening; chronic diseases, polypharmacy), in policy-driven research (e.g., mediation effect of health literacy in the relationship between socioeconomic status and health related outcomes, and in longitudinal study (e.g. identifying predictors of nursing home admission among older BHIS participants). The linkage of both data sources combines their strengths but does not overcome all weaknesses.

The availability of a national register number was an asset for HISlink. Policy-makers and researchers must take initiatives to find a better balance between the right to privacy of respondents and society's right to evidence-based information to improve health. Researchers should be aware that the procedures necessary to implement a link may have an impact on the timeliness of their research. Although some aspects of HISlink are specific to the Belgian context, we believe that some lessons learned are useful in an international context, especially for other European Union member states that collect similar data.

Keywords: record linkage; data linkage, health administrative insurance data, health claims data, health interview surveys

7.2. BACKGROUND

An evidence-based health policy requires sound and reliable health data and appropriate research methods from which it can be explored. To answer research questions, researchers can rely both on data derived from health surveys and on

administrative data, such as health insurance data, health care data from primary care or hospital information systems, disease-specific registers, etc. (1). Although administrative data is initially collected for other purposes, it is increasingly being used as a secondary data source for research. Such secondary data is generally easily accessible, resource-efficient and offers additional advantages, depending on the nature and the source (2).

Data linkage brings together information that relates to the same individual, family, place or event from different data sources (3,4). Single data sources are more commonly insufficient for answering complex research and policy questions. When answering these questions, the repeated collection of primary data is less flexible, more costly and more complex compared to data linkage. In countries where administrative data linkage is traditionally well established (e.g. in the UK, Australia, Canada, the Nordic countries, etc.), linked data is increasingly used for public health research purposes (5–7). Internationally, data linkage is common and an accepted practice for population health research and monitoring (8), especially to leverage existing data. Indeed, data linkage is a powerful and a cost-effective method for cohort studies. For example, in Germany, the lidA- leben in der Arbeit is a cohort study on work, age and health which uses survey data that is linked to claims data from a large amount of statutory health insurance data (9). Furthermore, such data linkage is a well-established method for external validations. Surveys data may be subject to bias (selection bias, recall bias) or may be inaccurate. Data linkage is a useful tool to validate such information. For instance, Hall et al. studied the validity of self-reported screening for prostate cancer and colorectal cancer in the United States (10). Van der Heyden et al. (2016) also assessed the validity of self-reported information on health care use (11). In another study, the same author estimated the predictive validity of the Global Activity Limitation Indicator (GALI) in the general population in Belgium (12).

In Belgium, the Belgian Health Interview Survey (BHIS) and the Belgian Compulsory Health Insurance (BCHI) are important sources of information on population health and healthcare consumption and are complementary. The National Institute for Health and Disability Insurance (NIHDI) commissioned a linkage study between BHIS and BCHI data with 3 specific questions: 1) to explore regional differences in healthcare consumption in more depth; 2) to assess the validity of

healthcare-consumption-based chronic disease indicators; 3) to estimate the cost to Belgian health insurance if some groups of non-reimbursed medicines (analgesics, laxatives and calcium supplements) were to be reimbursed (13). Moreover, the linked data was used in further studies (11,12). The HISlink project was then launched in 2017 as a systematic linkage between each wave of BHIS and BCHI data.

Linking BHIS and BCHI data sources allows the strengths of different data sources to be used synergistically and provides opportunities for new and advanced research. While BHIS data on medical consumption may be subject to recall bias, may be inaccurate and are prone to substitution by BCHI data, it is a source for detailed information on sociodemographic data, health-related behaviour and mental health. BCHI data also addresses elements that cannot be collected by means of a survey (e.g., healthcare expenditure, medical procedures).

While linkage of administrative-to-administrative data has a long tradition (9,14–19), linkage of survey data with administrative data is a relatively new field with great potential (9) and with its own challenges and considerations to take into account. These challenges may vary according to the context and the applicable data protection requirements. However, there is a paucity of information on the research opportunities and challenges faced when linking survey and administrative data. This study aims to fill these gaps.

Within the framework of HISlink, data from two BHIS waves has been linked to BCHI data: the BHIS2013 and BHIS2018. Using the case of these two linkages, this paper aims to discuss the methodology and the lessons on barriers and opportunities of linking survey data with health insurance data. More specifically, the focus will be on the following items: the practical implementation and outcomes in terms of linked datasets and the studies conducted, lessons learned and recommendations for future linkages. Although the Belgian context may be different from those of other countries, we believe that such information could be relevant for future researchers who plan to link surveys and health insurance data.

7.3. THE IMPLEMENTATION OF INDIVIDUAL DATA LINKAGE: AN EXPERIENCE BASED ON THE HISLINK STUDY

In Belgium, the BHIS and the BCHI have been linked for the last three waves of the BHIS, conducted in 2008 (as part of a feasibility study), 2013 and 2018. At the time of writing, the BHIS2008 data link had been destroyed due to the expiry of the retention period. Therefore, in this study, only the linkage of BHIS2013 and BHIS2018 are considered. This section describes the data sources, the linkage process and the privacy issues that arose and how they were overcome.

Description of data sources

HISlink combines BHIS and BCHI data sources. An overview of the most essential features of HISlink database is displayed in Table S1 (Supplementary file).

BHIS data

The BHIS is a national, cross-sectional household survey conducted every 5 years since 1997 by Sciensano, the Belgian health institute, among a representative sample of Belgian residents, including older, institutionalized people. Participants are selected from the national population register, using a multistage, stratified-sampling design (20). The participation rate of the survey at a household level was 57.1% and 57.5% for BHIS2013 and BHIS2018 respectively. Information is collected through a Computer-Assisted Personal Interview (CAPI) and a paper and pencil questionnaire for the more sensitive questions. Detailed methodology of the survey can be found in Demarest et al. (2013) (20). Though BHIS has several advantages: data are collected at a total population level, including people who do not make use of health services. Information is obtained from the perspective of the individual him/herself. The collection of self-perceived health, lifestyle and behaviour data is only (or mainly) possible through a survey. Information is collected simultaneously on individuals' health status, health behaviour and the use of health care, but also on socio-demographic health determinants, such as socio-economic status. This horizontal data collection makes it possible to study the relationship between different domains and topics. The different waves of BHIS enables trends analysis (21). However, as with all surveys, the organization of BHIS is expensive and time-consuming. Moreover, BHIS data is self-reported and therefore subject to biases such as selection

bias, recall bias or social-desirability bias (9,11,22,23). For instance, BHIS data on medical consumption may be subject to recall bias, may be inaccurate and prone to substitution by BCHI data (i.e. objective health-consumption data).

BCHI data

In Belgium, there is compulsory health insurance which is a source of exhaustive and detailed data on the reimbursed health expenses of almost 99% of the total population. However, there are some differences in coverage rates between regions and demographic characteristics (24). Since 2002, the InterMutualistic Agency (IMA), an overarching national organisation, collects and manages data on all Belgian citizens from these sickness funds (hereinafter referred to as BCHI data). The BCHI database is a longitudinal linkage between 3 components: the individual's background information, health-consumption data and database on use of outpatient medicines, which are linked using a Trusted Third Party (TTP) (25,26), i.e. the linkage was outsourced to another organisation that has access to identifiable data and has performed the linkage. The database includes an arbitrary id-code, allocated by the TTP. The primary goal of the BCHI data is for reimbursement purposes. BCHI data is widely used by important actors in the health field for reimbursement-related studies, assessment and planning of health care costs. In addition, BCHI data is also used for specific studies beyond its initial intended use (secondary use). One advantage is that the data is not self-reported, nor is it limited to a certain registration period, since there is continuous data collection for administrative purposes. Although BCHI data does not include information on the diagnosis, algorithms have been developed to estimate the prevalence of certain chronic diseases at a general-population level (pseudo pathologies derived from medication use) (27). Furthermore, this enables trend analyses and longitudinal studies (28,29). BCHI data has some shortcomings: the main limitation of the BCHI is that it only includes information on covered health services and goods, and there is a limited information on outpatient supplements. Next, since the purpose of BCHI data is the billing of services, the data may be subject to errors (e.g. inaccurate procedure codes, upcoding errors, duplicate billing) (30). Detailed information on BCHI data can be found elsewhere (31).

The above description shows that some information is only available in the BHIS (e.g. health status, health behaviour), other information is common to both data sources, even if conceptually different (e.g. health care utilisation, use of medication, as well

as a limited amount of socio-demographic information), while other information is only available in the administrative database (specific procedure codes such as nursing home admission, healthcare costs), which therefore makes the two databases complementary. The HISlink 2013 and 2018 resulted in datasets containing around 1200 variables and related indicators from BHIS and 130 variables or indicators from BCHI. Table S1 in supplementary file presents an overview of the content of the linked database, organised by modules, i.e. a set of information related to the same topic.

The partners involved, the linkage process and data flow

Figure 7.1 presents the data flow and the partners involved at each step. BHIS data is linked at an individual level to BCHI data, using the unique identifier: the national register number (NRN). The linkage is initially done by the reference person. At a later stage, household composition was compared according to BHIS and BCHI information and (based on date of birth, sex and date of the interview) the other household members' NRN were retrieved. The linkage process is quite complex since it requires several coding processes to ensure privacy and data protection. Detailed information on the linkage process and data flow is provided elsewhere (32). For the sake of clarity, the linkage scheme has been altered slightly. In summary, during the process, encrypted data are exchanged between the partners in a secure manner. For privacy reasons, there is need to ensure that none of the involved parties would have access to both the sensitive data and the NRNs during the linkage procedure. A small cell risk analysis (SCRA) is carried out by IMA. Only pseudomised data sets are then made available to Sciensano researchers on IMA server. Researchers have access to linked database through a Virtual Private Network (VPN) connection with secure token. Ultimately, a quadruple coding system ensures a coded database where no single party holds all of the respective keys enabling identification of individual patients.

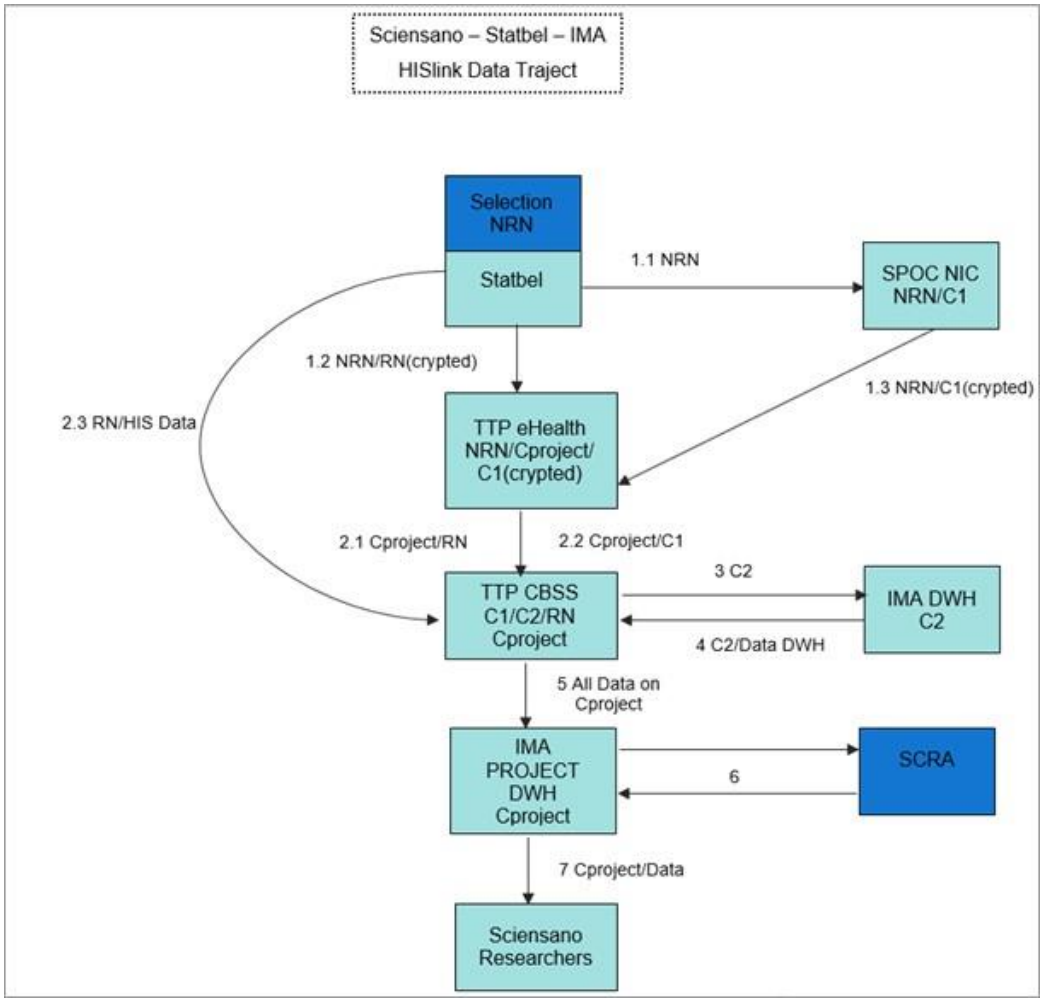


Figure 7.1: Step-by-step overview of linkage procedure and data coding system to enable data linkage for the HISlink 2018, Belgium (See legend below).

Legend of Figure 7.1

1. *NRN: National Register Number; Statbel: the Belgian statistical office; RN: Random Number; SPOC NIC: Single Point of Contact National InterMutualistic College; TTP CBSS: Trusted Third Party Crossroads Bank for Social Security; IMA DWH: InterMutualist Agency Data Warehouse; TTP eHealth: Trusted Third Party eHealth; SCRA: Small Cells Risk Analysis; C1/C2: coding 1/2; Cproject: project specific coding.*
2. *Explanatory note: the link involved the following steps, Figure 7.1.*
3. *Statbel selects the NRN of BHIS participants and transmits this selection of NRN to the NIC (1.1) and the selection of NRN with an internal RN (Random Number) specific to this project to the TTP eHealth (1.2). The NIC Security Advisor transmits an NRN/C1-encoded list of persons to the TTP eHealth, with C1 encrypted (1.3).*
4. *On the basis of a second coding (C1 → C2), the data are selected from the IMA DWH (3).*
5. *The data is sent back on a C2 basis to the TTP CBSS (4).*
6. *TTP CBSS replaces C2 with Cproject and also converts the received data into Cproject. These are transmitted to the IMA DWH (5).*
7. *A small cell risk analysis (SCRA) is carried out by the IMA (6).*
8. *The data sets are made available to Sciensano researchers (Cproject) (7).*

According to the GDPR, the processing of sensitive personal data, such as data concerning health shall be prohibited. However, processing for research is included as one of the exemptions of this rule under certains conditions. Article 5 of the GDPR defines some basic principles that must be taken into account when processing personal data (lawfulness, proportionality, accuracy, data minimization, storage limitation and integrity and confidentiality). The principle of proportionality means that researchers may only process personal data for the purpose of their research, and the processing must be reasonable and proportionate to the purpose of the research. Therefore, proportionality requires data minimisation, meaning that only that personal data which is adequate and relevant for the purposes of the processing is collected and processed (33,34).

Because of the proportionality principle, only a select amount of information from BHIS and BCHI data is included in the HISlink. An overview of BHIS and BCHI data included in the HISlink can be found in Table S1 (supplementary file). BCHI data covering the period from 2012 (or from 2008 in some specific cases such as dental care or cancer screening) to 2018 (or HISlink2013); and covering the period from 2017 (or from 2013 in some specific cases such as dental care or cancer screening) to 2023 (or HISlink 2018) is included in this study.

Privacy procedures

The BHIS2013 and BHIS2018 were carried out in line with the Belgian privacy legislation and have been approved by the Ghent University Hospital ethics committee on October 1, 2012 (opinion EC UZG 2012/658) and December 21, 2017 (opinion EC UZG 2017/1454) respectively. Participation in the BHIS is voluntary. No written consent was foreseen. Participation was equivalent to giving consent.

For the linkage to BCHI data, authorization was obtained from the Belgian Information security committee acting as an institutional review board (IRB) (local reference: Deliberation No. 17/119 of December 19, 2017, amended on September 3, 2019, for the HISlink 2013 and local reference: Deliberation No. 20/204 of November 3, 2020 for the HISlink 2018). In its deliberation, the IRB required Sciensano to inform the BHIS participants about the linkage of their data. In view of the disproportionate effort this would require (almost 11,000 individuals for the BHIS2013 and more than 12,000 individuals for the BHIS2018), and since the linkage process was launched before the

implementation of the GDPR, Sciensano presented an alternative approach to the IRB, which was accepted. This approach consisted of an exemption from obligation to provide information at an individual level, as well as communication about the data processing, provided to the general public, through a publication on the BHIS website.

Study population, linkage rates and an evaluation of linkage quality

All BHIS participants were eligible for inclusion in the HISlink. Figure 7.2a and Figure 7.2b present the selection process for the final participants: BHIS2013 and BHIS2018, respectively. Overall, the linkage rate was 92.3% for BHIS2013 and 94.2% for BHIS2018.

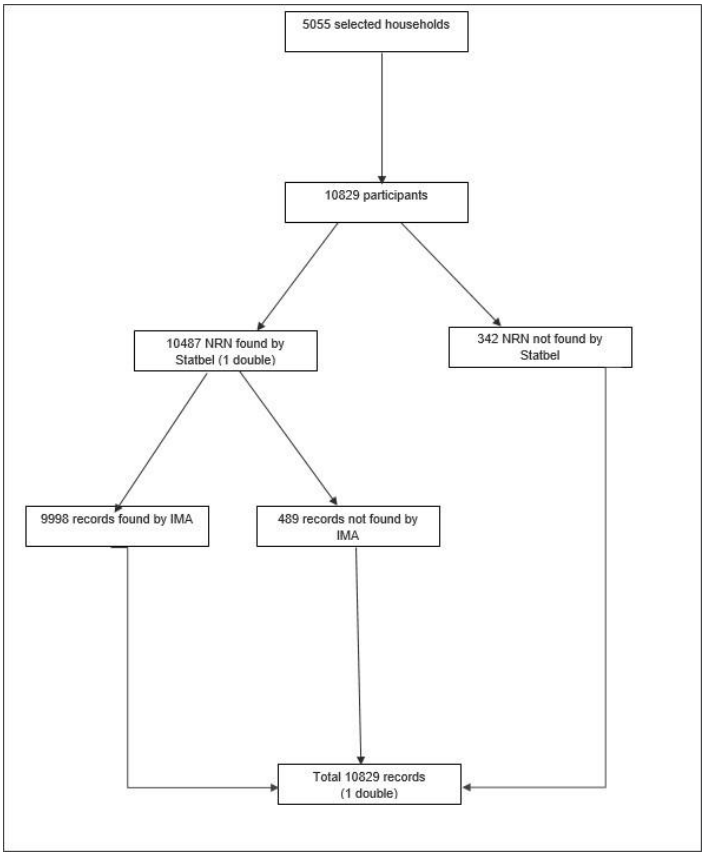


Figure 7.2a: Data flow and linkage global results, HISlink 2013, Belgium. *IMA: InterMutualist Agency, NRN: National Register Number*

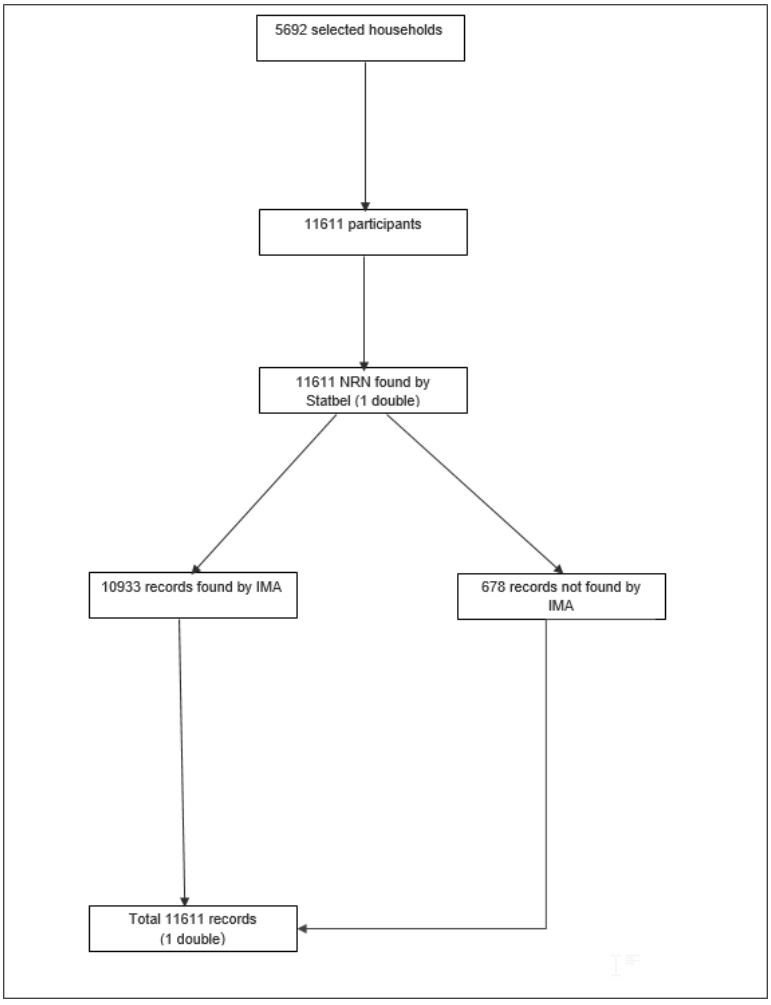


Figure 7.2b: Data flow and linkage global results, HISlink 2018, Belgium. *IMA: InterMutualist Agency, NRN: National Register Number*

Table 7.1 presents the linkage rates and the results of the evaluation of linkage quality. The linkage rates differed between population subgroups.

To assess the linkage quality, a comparison was made between the characteristics of linked and unlinked data (17,35). Standardized differences of the proportions were used to test for statistically-meaningful differences between those with linked and those with unlinked data (36–38). The standardized difference was the difference in the two proportions, divided by an estimate of the prevalence of the covariate in each of the two groups (37). A value equal to or greater than 0.10 was considered

significant (37,38). Significant differences were observed between respondents with linked and unlinked records in terms of age, educational attainment, household composition, nationality, household income and the region of residence both for HISlink 2013 and HISlink 2018, while significant differences were observed according to gender, with a lower linkage rate for males, for HISlink 2018 only (Table 7.1).

Table 7.1: Characteristics of the study population with linked and unlinked data, HISlink 2013 and 2018, Belgium

	HISlink 2013				HISlink 2018			
	Linked N=9998	**Unlinked N=831**	**Standardized difference**	**Linkage rate (%)**	**Linked N=10933**	**Unlinked N=678**	**Standardized difference**	**Linkage rate (%)**
Characteristics								
Gender, n (%)								
Male	4819 (48.8)	412 (47.4)	0.03	92.1	5235 (49.0)	353 (56.7)	-0.16	93.7
Female	5179 (51.2)	419 (52.6)	-0.03	92.5	5698 (51.0)	325 (43.3)	0.16	94.6
Age, n (%)								
0-14	1523 (17.2)	193 (27.0)	-0.24	88.7	1766 (17.7)	92 (13.6)	0.11	95.0
15-24	1051 (11.5)	100 (13.1)	-0.05	91.3	994 (11.3)	65 (11.1)	0.01	93.8
25-34	1272 (12.3)	134 (15.3)	-0.08	90.5	1254 (12.8)	84 (15.7)	-0.08	93.7
35-44	1378 (13.7)	144 (14.8)	-0.03	90.5	1461 (12.7)	117 (12.9)	-0.01	92.6
45-54	1445 (14.9)	113 (13.3)	0.05	92.7	1569 (13.8)	156 (21.2)	-0.19	90.9
55-64	1379 (12.5)	71 (7.5)	0.17	95.1	1584 (13.1)	86 (14.8)	-0.05	94.8
65-74	998 (9.0)	34 (3.7)	0.21	96.7	1249 (9.7)	40 (5.4)	0.16	96.9
75+	952 (8.9)	42 (5.3)	0.14	95.8	1056 (8.9)	38 (5.3)	0.14	96.5
Education, n (%)								
Primary/No diploma	1054 (9.4)	79 (9.1)	0.01	93.0	779 (5.8)	32 (5.1)	0.03	96.0
Lower secondary	1389 (12.3)	64 (8.9)	0.11	95.6	1391 (12.2)	43 (6.9)	0.18	97.0
Upper secondary	3194 (33.3)	201 (24.6)	0.19	94.1	3279 (32.0)	123 (18.9)	0.30	96.4
Higher education	4211 (43.8)	468 (55.7)	-0.24	90.0	5309 (48.8)	446 (67.3)	-0.38	92.2
Missing	150 (1.2)	19 (1.7)	-0.05	88.7	175 (1.2)	34 (1.8)	-0.04	83.7
Household composition, n (%)								
Single	1685 (15.1)	78 (10.9)	0.12	95.6	2047 (15.5)	104 (14.1)	0.04	95.2
One parent with child(ren)	1115 (9.0)	87 (7.8)	0.04	92.8	1228 (10.9)	48 (6.6)	0.15	96.2
Couple without child(ren)	2203 (22.1)	125 (17.6)	0.11	94.6	2469 (22.4)	129 (24.2)	-0.04	95.0
Couple with child(ren)	4105 (45.2)	374 (45.9)	-0.01	91.6	4656 (46.3)	361 (51.5)	-0.11	92.8
Other or unknown	890 (8.6)	167 (17.8)	-0.27	84.2	533 (4.9)	36 (3.6)	0.07	93.7

Nationality, *n (%)*								
Belgian	8834 (91.4)	457 (60.0)	0.79	95.1	9461 (90.1)	300 (50.1)	0.97	96.9
Non Belgian - EU	700 (4.9)	276 (29.1)	-0.68	71.7	846 (5.2)	338 (43.0)	-0.98	71.4
Non-Belgian - non EU	457 (3.6)	98 (10.9)	-0.28	82.3	621 (4.7)	40 (6.9)	-0.09	93.9
Missing	7 (0.1)	0 (-)	-	100.	5 (0.1)	0 (-)	-	100
Household income, *n (%)*								
Quintile 1	1983 (16.9)	141 (21.3)	-0.11	93.4	1192 (8.9)	29 (6.0)	0.11	97.6
Quintile 2	1516 (14.9)	57 (10.9)	0.12	96.4	1450 (11.9)	26 (3.2)	0.33	98.2
Quintile 3	1748 (18.7)	93 (13.4)	0.14	94.9	1820 (16.5)	41 (8.3)	0.25	97.8
Quintile 4	1768 (20.5)	83 (12.8)	0.21	95.5	2322 (22.4)	84 (11.8)	0.28	96.5
Quintile 5	1781 (19.7)	193 (22.1)	-0.06	90.2	2487 (26.4)	317 (47.5)	-0.45	88.7
Missing	1202 (9.3)	264 (19.4)	-0.29	82.0	1662 (13.9)	181 (23.2)	-0.24	90.2
Region of residence, n (%)								
Flanders	3425 (57.9)	87 (32.0)	0.54	97.5	4230 (56.5)	66 (31.4)	0.52	98.5
Brussels	2715 (10.2)	388 (29.6)	-0.50	87.5	2873 (10.2)	226 (26.2)	-0.42	92.7
Wallonia	3858 (31.9)	356 (38.4)	-0.14	91.5	3830 (33.3)	386 (42.4)	-0.19	90.8

7.4. OUTCOMES OF LINKED DATA - ADDED VALUES OF HISLINK FOR EPIDEMIOLOGICAL RESEARCH

Linking BHIS to BCHI data has resulted in a richer database, which has allowed studies to be carried out that would not have been possible using the two sources separately. Table 7.2 gives some examples of studies undertaken using the linked database. These examples illustrate the added value of the HISlink data for public health research. The studies carried out can be grouped in terms of different benefits or objectives in validation studies, policy-driven research studies and longitudinal studies.

Validation studies

The linked data offered opportunities to answer methodological questions on the validation of survey information, such as the validity of self-reporting or conversely on the validation of administrative information. For instance, data on the mammography uptake is usually based on self-reports in population-based surveys such as BHIS. However, the validity of self-reported information through surveys is a concern, due to the associated potential reporting bias. To gain further insights into the validity of self-reported breast cancer screening in Belgium, we assessed the selection and reporting biases of BHIS-based estimates in the target group (women aged 50–69 years) using reimbursement data for mammograms taken from the BCHI. We found that the validity of self-reported mammogram uptake in women aged 50–69 years, is affected by both a selection and reporting bias (overreporting) and caution should therefore be exercised when using BHIS information as the sole source for assessing mammogram uptake (22).

Currently, the estimation of the prevalence of many chronic diseases in Belgium is still often based on self-reported BHIS data. On the NIHDI's initiative, we evaluated whether BCHI data can be used to ascertain the prevalence of chronic diseases in the Belgian population. For this purpose, we assessed the agreement between the definitions used in health-administration cases (algorithms based on Anatomical Therapeutic Chemical (ATC) codes of disease-specific medication) and the definitions used in self-reported cases (based on the response to the following question: *"Have you suffered from any of the following diseases in the last 12 months?:"* diabetes, asthma, chronic obstructive pulmonary disease (COPD), cardiovascular diseases

including hypertension (CVDs), Parkinson's disease, thyroid disorders and epilepsy in the Belgian population. We concluded that BCHI's chronic-disease case definitions are an acceptable alternative for identifying cases of diabetes, CVDs (including hypertension), Parkinson's disease and thyroid disorders, but yield a significantly-underestimated number of patients suffering from asthma and COPD (27).

Another study explored the differences between self-reported and prescription-based estimates in the prevalence and determinants of polypharmacy in the older, general population in Belgium; and assessed the relative merits of each data source. The key findings were that surveys and prescription data measures polypharmacy from a different perspective, but overall conclusions in terms of prevalence and determinants of polypharmacy do not differ substantially according to data source (30).

Policy-driven research

The linked database served as an evaluation tool for policy measures. Indeed, in our study "Effectiveness of protective measures on dental care use: analysis from linked database" we assessed the effectiveness of financial protective measures on the use of dental care among a representative sample of Belgian adults. We concluded that the current health interventions in dental care use are not yet effective for vulnerable people (39).

The reduction of socioeconomic (SE) health inequalities is an important objective for public health policies. It is therefore important to identify factors that contribute to these inequalities. Health literacy (HL) is of interest as it constitutes a potential pathway by which socioeconomic status (SES) affects health. In contrast to a number of socioeconomic factors that are more difficult to modify, HL is a more easily modifiable factor. As such, HL can also be taken into account in the attempt to reduce health inequalities. If HL is an important mediator in explaining SE health differences, actions to improve HL in low SE groups will reduce SE inequalities. This study explored whether HL acts as a mediator in the association between SES as measured by educational attainment and household income and a selected health (-related) outcomes that were of great interest from public health perspective in various domains: (1) health behaviour (physical activity, type of diet, alcohol and tobacco consumption), (2) perceived health status (self-rated health (SRH)), (3) use of curative care (purchase of antibiotics and antidepressants), and (4) use of preventive care

(preventive dental care, influenza vaccination, breast cancer screening). The study showed that HL partially mediated the relationship between education and health behaviour (except tobacco consumption), perceived health status, purchase of antidepressants and preventive dental care, accounting for 4.4% to 15.4% of the total effect. As far as the association between household income and health (-related) outcomes is concerned, the findings showed that HL constituted a pathway by which household income influences health behaviour (except alcohol consumption), perceived health status, purchase of antidepressants) and preventive dental care, with the mediation effects accounting for 4.2% to 12.0% of the total effect (40).

The linked data has been used to estimate the annual costs in health care and lost productivity associated with excess weight among the adult population in Belgium. The study concluded that BMI is a substantial social-economic burden in Belgium. Every year at least €4.5 billion are spent to cover the direct and indirect costs related to excess weight and obesity. Policies and interventions are urgently needed to reduce the prevalence of excess weight and obesity, thereby decreasing these substantial costs (41).

Longitudinal study

The linked data not only increases the number of variables. By following-up on BHIS participants up to 5 years after the survey, research questions can be addressed that require a longitudinal design. In this context, we estimated the risk of nursing home admission (NHA) among the older population of 65+ years and its predictors in Belgium. We found that the cumulative risk of NHA was 1.4%, 5.7% and 13.1% at, respectively 1 year, 3 years and 5 years of follow-up. A higher age, living arrangements, falls, physical chronic conditions and mental disorders such as Alzheimer's disease, appeared as strong predictors of NHA (29).

The HISlink data was further used to investigate the association between polypharmacy and mortality in the community-dwelling older population. It was found that polypharmacy affects the mortality of older people who are in relatively good health and concluded that a critical evaluation of polypharmacy in older people aged below 80 years and in people without severe functional limitations may reduce mortality in these population groups (42).

Table 7.2: Examples of epidemiologic research with HISlink data, Belgium

Study	Research questions	Key findings	Reference
Validity of mammography uptake in women aged 50-69 years	To which extent self-reported mammography uptake from BHIS is valid as compared to objective information from IMA?	The validity of self-reported mammography uptake in women aged 50-69 years is affected by both selection and reporting bias. Cautiousness is needed when using self-reported estimates as the sole method to quantify mammography coverage.	Berete et al. (2020) (22)
Ascertainment of chronic diseases	Can the indicators of pseudopathologies in administrative data (IMA data) be used to assess prevalence of chronic diseases in the general population?	The indicators of pseudopathologies are an acceptable alternative to identify cases of diabetes, CVDs, Parkinson's disease and thyroid disorders but yield in a significant underestimated number of patients suffering from asthma and COPD. Further research is needed to refine the definitions of CDs from administrative data.	Berete et al. (2020) (27)
Impact of financial protective measures on dental health care use	What is the effectiveness of financial protective measures on the use of dental care among a representative sample of Belgian adults?	Current health interventions are not yet effective for vulnerable people in dental care use. High expenses as a result of chronic diseases are not associated with more postponement of dental care. More targeted financial interventions should be necessary to reduce postponement of dental service utilization.	Berete et al. (2020) (39)
Nursing home admission in older population in Belgium	What is the risk of nursing home admission among older population of 65+ years in Belgium? What are the predictors?	The cumulative risk of NHA was 1.4%, 5.7% and 13.1% at, respectively 1 year, 3 years and 5 years of follow-up Higher age, living arrangements, use of home care services, falls, urinary incontinence, subjective health, limitations, depression, Alzheimer disease, etc., appeared as strong predictors nursing home admission.	(Berete et al., 2022) (29)
Mediation effects of health literacy	Does health literacy mediate the relationship between socioeconomic status and health related outcomes in the Belgian adult population?	HL partially mediated the relationship between education and health behaviour (except tobacco consumption), perceived health status, purchase of antidepressants and preventive dental care, accounting for 4.4% to 15.4% of the total effect. Health literacy also mediated the association between income health behaviour (except alcohol consumption), perceived health status, purchase of antidepressants and preventive dental care, with the mediation effects accounting for 4.2% to 12.0% of the total effect.	Berete et al. Will be submitted to BMC Public Health (40)
Assessing prevalence of polypharmacy	What is the differences in the prevalence and	Surveys and prescription data measure polypharmacy from a different perspective, but overall conclusions in terms of prevalence and determinants of	Van der Heyden et al. (2021) (30)

among older adults	determinants of polypharmacy in the older population between self-reported and prescription based estimates? What is the relative merits of each data source?	polypharmacy do not differ substantially by data source.	
Association between polypharmacy and mortality in older population	What is the association between polypharmacy and mortality in the community dwelling population of 65+ years in Belgium?	Polypharmacy affects the mortality of older people in relatively good health. A critical evaluation of polypharmacy in older people below 80 years and in people without severe functional limitations may reduce mortality in these population groups.	Van der Heyden et al. (2021) (42)
Costs associated with excess weight in Belgium	What are the annual health care and lost productivity costs associated with excess weight among the adult population in Belgium, using national health data?	BMI has a substantial societal economic burden in Belgium. Every year at least €4.5 billion are spent to cover the direct and indirect costs related to overweight and obesity. Policies and interventions are urgently needed to reduce the prevalence of overweight and obesity thereby decreasing these substantial costs.	Gorasso at al. (41)

7.5. LESSONS LEARNED AND RECOMMENDATIONS FOR FUTURE LINKAGES

Although linking survey data with administrative data opens new research opportunities as presented above, such linkage is not without challenges. This section describes the main challenges and considerations that may be encountered in data linkage processes and a number of recommendations for future linkages will be formulated. Table 7.3 provides a summary of the challenges and considerations and the corresponding recommendations.

7.5.1. Lessons learned from to the linkage processes overall

Technical and operational issues of the linkage

The technical challenges inherent in linking survey data with administrative data are mainly related to the data quality and to the linkage errors (43). Next to these issues, the proportionality principle, infrastructure and statistical challenges are also important.

The quality of the data sources, i.e., the availability, completeness and discriminatory power of identifiers or key personal variables that can be used to construct the linkage key, is very important and determines the choice of linkage methods.

In some countries, a unique personal number, such as the NRN in Belgium or the personal identity number in Scandinavia, is required for access to almost all administrative services, including healthcare services use for each resident and can be readily used to obtain information about individuals. Such identifiers allow the linkage to be relatively straightforward (deterministic linkage approach), and make it possible to link data from many different administrative sources with marginal error (44). With regards to HISlink, the use of the NRN as a linkage key was a great asset. Moreover, such a unique identifier increases the linkage rate, although this rate varies between subgroups as shown in Table 7.1. About 8% of the BHIS2013 and 6% of the BHIS2018 could not be linked. This result could be explained by the fact that the BHIS household composition can deviate from the “official” household composition in the national register, preventing the linkage. In addition, as Table 7.1 shows, the linkage was not possible for a number of people who are more likely to be from the Brussels-

Capital Region and more likely to be EU nationals. This sub-group could probably be people working for EU institutions, other international organizations or posted workers from other EU countries, living and working in Belgium but insured in their country of origin. Therefore no data could be retrieved fro:m the BCHI.

In many other countries however, unique identifiers are not available and this might constitute an important barrier to linking the same person across multiple data sources (18). In such contexts, linkage often depends on the use of non-unique 'imperfect' identifiers such as name, postcode, date of birth or other indirect identifiers. In combination, these variables can make it possible to identify records that belong to the same person, using more complex algorithms (probabilistic linkage approach). The probabilistic linkage method is the most common approach, usually in combination with the deterministic methods (45,46).

The second challenge when linking survey data to administrative data is the risk of linkage errors, which typically occur where there is no unique identifier across different data sources (47) or in the event of imperfect identifiers. This problem could result in substantially biased results (17,48). Linkage errors arise when pairs of records are incorrectly classified. False-matches occur when records from different individuals link erroneously, while missed-matches occur when records from the same individual fail to link (45,46). Data analysts should therefore evaluate the quality of linked data by measuring linkage errors before proceeding with any further analysis. The availability of similar information in both data sources or in a reference database will be helpful in this regard. For HISlink, comparing age, sex, region of residence and the prevalence of certain chronic diseases, we detected an error in the previous version of HISlink 2018 data due to the use of the wrong database during the linkage process. This error was corrected by the linkage TTPs afterwards.

Another challenge that researchers face in data linkage is the proportionality principle, which means that only those variables that are relevant to the purpose of the study should be selected to avoid the re-identification of individuals. In this context, researchers should have a thorough knowledge of their data sources. The selection of relevant variables must be done precisely before the linkage process. The more information there is in both data sources, the more difficult this task becomes. However, this approach is not optimal as it is time-consuming and requires an in-depth knowledge of the data sources. In addition, when it is necessary to include new

relevant variables or indicators that have been forgotten, the whole process has to be restarted (new IRB opinion, new linkage, etc.). An alternative, perhaps better approach could be too ask for permission to link both datasets completely in a first step. In a second step, each research project demands in a simplified procedure access to the relevant variables of the fully linked dataset in accordance with the proportionality principle. This is basically what is done at Statistics Netherlands (49– 51).

Further consideration for researchers wishing to link data is the infrastructure needed to store and access the linked data. Some questions need to be answered beforehand: how will data be stored safely? What is the cost for the infrastructure? How will data be protected? How can data be accessed in a safe and easy way? (28). In the case of HISlink, the linked data was stored on the IMA server and researchers access it securely using a token.

Finally, analysing linked datasets raises a number of additional 'statistical' challenges for researchers. Although linked data has several advantages, it is important to bear in mind that the limitations of both data sources remain even after the linkage. Researchers need to be aware of this to understand and interpret the results carefully. In addition, in the event of linkage errors, specific statistical methods need to be applied (35,46). Furthermore, with the complexity of administrative data, it is often necessary to involve an expert on this data in the analysis stages as well as when interpreting the results. In our case, the BCHI data is collected for administrative purposes, not for epidemiological research. It is therefore not easy to understand and use. Expert advice is often needed to make good choices when planning the analysis. The IMA's single point of contact and the many experienced Sciensano researchers are well-qualified to fulfil this requirement.

Ethical, legal and societal aspects

The most important concerns facing data linkage are privacy and confidentiality issues (52). With the implementation of the GDPR in 2018, new decision-making bodies were established for the authorisation of data linkage, and privacy and confidentiality issues were redefined. Because of these confidentiality issues, institutional review board (IRB) approval is often required to link the data. However, such IRB approval processes are usually complex and time-consuming, especially when the linkage is

not consent-based. For both HISlink 2013 and HISlink 2018, it took several months to get the IRB approval. Therefore, to facilitate data linkage and overcome the lengthy negotiation and ad hoc approval processes for each BHIS-BCHI linkage, it would be useful to set up some kind of umbrella agreement protocol for public institutions such as Sciensano, to cover several years and several waves of BHIS-BCHI linkages.

To preserve privacy and prevent the disclosure of sensitive information, data linkage often relies on the separation principle of linkage and analysis processes, meaning that those conducting the linkage (often TTPs) only have access to a set of identifiers, whilst those analysing the linked data only have access to de-identified attribute data (17). However, this type of approach causes a significant delay in the linkage process due to the administrative steps that take time (e.g. the signature of an official agreement between the parties involved). Furthermore, although this approach reduces the risk of disclosure of sensitive information about individuals, it means that important aspects of the linkage process are obscured, which makes it difficult for researchers to judge the reliability of the resulting linked data for their required purposes (17,47).

Respecting respondents' rights and maintaining their trust are further considerations. According to the new EU data Act, trust and altruism are essential in secondary data use (53). When researchers plan to link data as part of a future survey, citizens must be able to decide whether they want to share their data, they must be informed that their data is being used and by whom. In other words, they need to opt-in through informed consent (1, 9, 54, 55). Informed consent is required to ensure that respondents are aware of the risks and benefits involved in releasing and linking their personal data for research purposes, even though obtaining the opt-in linkage consent from all respondents is a challenging task. To link historical survey data to administrative data, there are exceptions to the requirement for informed consent, especially if contacting study participants is impossible or unreasonable (1, 9). The GDPR contains specific exemptions to informed consent as a legal basis for the use of data to escape a 'consent or anonymise approach' or a 'fetishisation of consent', especially in the case of observational health research (56). For the BHIS2013 and BHIS2018 linkages, because of the disproportionality to inform and seek consent from all BHIS participants and also because the authorization procedure was implemented prior to the GDPR, we proposed that the acquisition of consent from BHIS participants

was obtained by way of a waiver, and this approach was accepted by the IRB. While these exemptions to informed consent are possible for historical data linkages, for any planned future linkages, researchers must seek informed consent from participants during the survey.

7.5.2. Lessons learned related to the outcomes

Without a doubt, the HISlink offers the potential to obtain more comprehensive data on the population's health, facilitating new research perspectives for public health as demonstrated in this study. The BHIS data are only available every 5 years and some studies require more comprehensive data than the current linked data. The HISlink can be seen as a first step towards more comprehensive data linkages. To ensure that the benefits of data linkage are fully maximised, it is important to consider the inclusion of other administrative data such as hospital discharge data, mortality data, environmental data, primary electronic medical record (EMR), etc. For example, extending linked data to hospital discharge data could help target internal quality improvement efforts for specific patient groups (e.g., preventive care for diabetics) or help assess the determinants of hospitalisation and understand the underlying factors that influence length of hospitalisation. A linkage with the EMR may also be useful for studying appropriate polypharmacy, for example. However, in some countries such as Belgium, there is currently no integrated primary EMR. Only a few sentinel networks exist, such as the Intego database. For the future, consideration needs to be given to establishing a legal framework for such an integrated database.

At international level, the linkage between survey and administrative data has also proven its value. Indeed, such a linkage has been widely used in validation studies (10,57,58), but also in addressing specific research questions. For example, using health survey data linked to administrative health services data, the Institute for Clinical and Evaluative Sciences (ICES) researchers in Ontario, Canada, developed and validated an algorithm for population-based prediction of diabetes - the Diabetes Population Risk Tool (DPoRT) that accurately predicts diabetes risk in a population (59). The linkage of Canadian Community Health Survey (CCHS) with medical claim data, has been used to investigate individual-level characteristics that are associated with community-dwelling high-cost users. They found that high-cost users status was strongly associated with being older, having multiple chronic conditions, and reporting

poorer self-perceived health. The authors further found that high-cost users tended to be of lower socio-economic status, former daily smokers, physically inactive, current non-drinkers, and obese (60). Finally, the linkage of survey and administrative data has been used to address methodological issues such as bias adjustment (61–63) or non-response analysis (64).

The BCHI data does not contain clinical information. In addition, there is no information on non-reimbursed care in the BCHI data. Although information is available on vital status, there is no information on cause of death. The absence of such important information prevents some policy-oriented research questions from being answered better. In future, efforts could be made to include more data sources in HISlink, and an initial step would be to include hospital discharge data.

The BCHI data is only available two years after consumption, meaning that the linkage can only be made with a two-year delay which precludes 'real time' linkage. Data availability should be accelerated in the short to medium term given the widespread use of electronic billing.

Furthermore, with the limited sample size of the BHIS (about 10,000 participants), subgroup analysis is impossible or yields inaccurate results, for example for rare events or specific subgroups.

Finally, access to linked data is thus far highly restricted due to legal constraints. Only Sciensano researchers that are registered with the IMA as the users of the linked data have access to the data. To take further advantage of the linked data, the data owners, i.e., Sciensano, the IMA and the sponsor (NIHDI) could retain ownership but make the data available to other research studies in line with the primary objective of HISlink, subject to the owners' approval. One example of such an approach in cancer research is the National Cancer Institute's (NCI's) linked Surveillance, Epidemiology and End Results (SEER)-Medicare files where the NCI retains ownership of the data and releases it for approved research studies that guarantee the confidentiality of the patients and providers in the SEER areas (65).

7.5.3. Recommendations for future linkages

This study provides important information with regard to the individual linkage of survey data and health-insurance administrative data that other studies can build on.

Based on our experience, there are a number of aspects that need to be taken into account to ensure the success of data linkage in future research. The recommendations related to the ethical, legal and societal aspects, technical, practical challenges, as well as those related to the outcomes are summarized in Table 7.3, and the main ones are further elaborated below.

Recommendation 1: Gain and maintain the citizens trust in secondary use of data and data linkage

With the implementation of the GDPR, the consent form became mandatory for future planned linkages. Researchers need to put in place strategies to gain the trust of and to involve citizens whose data will be linked (66). The perceived risk to privacy and data confidentiality constitutes one of the primary reasons why respondents decline the linkage request (55). It is therefore important to emphasise the merits of the research, to stress the importance of altruism (contribution to society) and to address respondents' privacy and confidentiality concerns by informing them of the safeguards put in place to protect their data.

Recommendation 2: Improve the communication with the participant, so there is more willingness to give a consent for linkage

The literature suggests a strong correlation between respondents' understanding and how likely they are to give consent (55,67). To achieve higher consent rates, it is necessary to shed light on respondents' understanding of the linkage consent. Several approaches have been proposed to improve linkage consent rates. One of these consists of providing key subgroups that are less likely to understand the linkage request, with additional targeted explanatory or informative material. Another approach would be to use tailored messages by asking the consent understanding questions first, then doing a targeted intervention to address any misunderstandings, before administering the linkage request. It is preferable to ask for linkage consent upfront, which yields higher consent rates (9, 45, 50, 51).

Recommendation 3: Adapt the need for consent to the context of the linkages

For linkages between datasets that already exist, a clear framework of acceptable practices needs to be developed, which the European Health Dataspace initiative is attempting to do (70). To maintain population trust in secondary use of data and data linkage, it is imperative that this framework is in line with citizens' values (66). A clear distinction should be made between:

1) Routine linkages, which are usually for primary use and where implicit consent can be assumed because it concerns direct clinical care. However, a harmonized framework needs to be developed in order to streamline secure data flows;

2) Necessary linkages, in a public health crisis, as exemplified by the COVID-19 pandemic and where consent should not be required (71); and

3) Linkages for public health research and surveillance or other scientific research in the public interest, where the preferred legal basis should not be consent, but an explicit legal and ethical framework that is developed by the national health data authorities, resulting in a federated network of Findable, Accessible, Interoperable and Reusable (FAIR), linkable data sources governed by rules that are trusted both by researchers and citizens.

Recommendation 4: Advoid the 'link and destroy model'

Many challenges remain before this can become a reality, but it would resolve the administrative burden, the need for case-by-case consideration and the overall uncertainty and inefficiency surrounding data linkage (72). From a broader perspective, it will be useful to have streamlined approval processes for efficient data access. Indeed, some jurisdictions adopt approaches for timely and cost-effective access to linked data (e.g. those in Ontario, Wales and Australia where linkage keys can be held in perpetuity), others such as in Belgium are restricted by the 'link and destroy' model, where linked data cannot be reused or are destroyed after a predefined data-retention time. In turn, these impact on the availability and accessibility of data for research and policy development (17).

Recommendation 5: Take up initiatives to work towards a better balance between the right to privacy of respondents and society's right to evidence-based information to improve health

Privacy considerations must strike a balance between the privacy rights of respondents and society's right to evidence-based information to improve health.

Although the separation principle of linkage and analysis processes (as implemented at: the Data Linkage Branch in Western Australia, the Centre for Health Record Linkage (CHeReL) in New South Wales (73), the Secure Anonymous Information Linkage (SAIL) Databank in Wales (74), the Centre for Data Linkage (CDL) in Australia (75), the Manitoba Centre for Health Policy in Canada (73)) is recognised as

good practice for protecting confidentiality, allowing linkage and analysis to take place together provides opportunities for both in-depth evaluation of linkage quality, and methodological advances in linkage techniques (76,77). Such an approach is in operation at the Institute for Clinical Evaluative Sciences (ICES) in Ontario. The ICES is legally allowed to receive fully identifiable data in order to perform linkage, to assess data quality and to provide coded data to research staff within the organisation. They operate a hierarchical access policy, which means that only a specific number of people have the highest level of access to all data elements, and most researchers can only access de-identified, coded data relevant to their study (73). The linkage approach as applied at Statistics Netherlands constitutes a good practice in Europe (49–51).

Recommendation 6: Optimize the way to deal with ethical and privacy requirements in order to be able to carry out data linkages in a reasonable time.

Beside the privacy and confidentiality issues, researchers should be aware of some technical aspects such as the complexity of the linkage process which often results with a delay in the linkage process. Getting the agreement signed between the parties involved was a crucial factor in delaying the process, especially when several parties are involved. Therefore, a formal, pre-established accreditation that negates the need for new signatures at each linkage (ad hoc approval) for institutions that are entitled to request a data linkage, would be a further step towards reducing the delay and facilitating the data linkage process.

Recommendation 7: Plan ahead the linkage of survey and administrative data, particularly where there is no unique identifier that can be used as a linkage key

If the linkage cannot rely on a unique identifier, researchers should identify more relevant variables (e.g., age, gender, date of birth, name, etc.) that will allow the construction of an almost perfect identifier for probabilistic linkage. As data linkage often relies on the separation of linkage and analysis processes, researchers should assess the linkage errors and quality of the linked data before conducting any further analysis. Several methods can be used to evaluate linkage quality, including the use of gold standard or reference data, sensitivity analyses, a comparison of the characteristics of linked and unlinked data, or post-linkage data validation (17,35).

Recommendation 8: Apply strategies to improve the linkage rates

Although the use of deterministic linkage methods has resulted in a relatively higher linkage rate, this approach is known to give rise to a number of missed matches (e.g. in the case of even a single digit error in the NRN). Therefore, a combination with subsequent probabilistic methods for unlinked cases to the deterministic linkage step would certainly result in a higher linkage rate. In addition, another explanation why the linkage was not always possible for everyone would be that only the NRN of the reference person was available and the others had to be found on the basis of household composition and socio-demographic characteristics. This approach is probably linked to the BHIS sampling strategy. However, BHIS household composition may differ from BCHI household composition or may change over time. Therefore, including the NRN of all individuals included in the survey, regardless of household composition would probably improve the linkage.

Recommendation 9: Demonstrate to funders and policy makers the usefulness of linkages, raise awareness of such initiatives and continue to promote the linkage between databases

The linked data is an important source for population health research. Its use by researchers can bring huge benefits in terms of providing a more complete picture of the population's health. However, within the context of budgetary constraints, it is important for researchers to demonstrate to funders and policy makers the usefulness of such linkage in order to maintain project funding and sustainability and to raise awareness of such initiatives. From a public health perspective, policy makers should continue to invest in data linkages; and the inclusion of other data sources (such as primary-care data and hospital discharge data) will augment the use of the linked data to expand the evidence base for policy makers and practitioners, which could therefore enrich population-based surveillance and research in the field of public health. However, in that case, there is a need to develop an overarching infrastructure. Since making linkages between multiple datasets would be very challenging, to be really cost-effective, it would be better to have an infrastructure that would allow access to different research institutes.

Recommendation 10: Consider substituting HIS information by administrative data as much as appropriate

In view of the current challenges facing surveys, there is need to keep survey questionnaires as short as possible. Hence the more information can be obtained through other sources, the shorter can be the questionnaire. When possible, self-reported items should be replaced by administrative data. This will be the case, for example, for cancer screening, reimbursed healthcare use or reimbursed drug use. However, it is important to keep in mind that the replacement of self-reported information by administrative data can have certain limitations since administrative data have their own shortcomings (e.g., incomplete or missing data, recording errors).

Table 7.3 : Overview of challenges, considerations and recommendations in linking surveys data with administrative data; HISlink, Belgium

Category	Description	HISlink-specific experience	Recommendations
Technical, practical challenges			
Data quality	Availability, completeness and discriminatory power of identifiers	National register number available and used as linkage key	Use unique identifier when available. Otherwise, carefully select linkage variables to construct linkage the keys. Ensure that these variables are as complete as possible (less missing values, less errors) and that no duplicate records exist in each data source.
Linkage errors	Usually arises in data linkage, typically when 'imperfect identifiers' are used and could result in substantially biased results. False matches (i.e., when records from different individuals link erroneously) and missed matches (i.e., when records from the same individual fail to link) (45,46) are of greatest concern. The number of false matches and missed matches can directly affect the estimation of prevalence or incidence rates. False matches (low specificity) lead to overestimates of prevalence whilst missed matches (low sensitivity) lead to underestimates. The impact of linkage error depends on the underlying prevalence of the target condition: analyses of rare conditions are more severely affected by linkage error compared with more common conditions, as overestimation is inversely related to the underlying prevalence (46).	Negligible/marginal false matches because of the accuracy of the linkage key. However, up to 8% of missed matches (see section 4.1 for possible explanations). The comparison of linked and unlinked records identified subgroups that are more prone to linkage errors (see Table 7.1).	Evaluate linkage quality and assess the impact of linkage errors on the results (17,35,46). The evaluation of linkage quality is vital to producing reliable results from studies using the linked data. Several methods can be used to assess linkage quality and errors: - comparing linked data with reference or 'gold-standard' datasets where the true match status is known; - structured sensitivity analyses where a number of linked datasets are produced using different linkage criteria; - comparisons of characteristics of linked and unlinked data to identify any potential sources of bias; - statistical methods accounting for linkage uncertainty within analysis (e.g. using missing data methods); - quality control checks (implausible scenarios) - sensitivity (proportion of matches that are correctly identified as links), specificity (proportion of non matches that are correctly identified as non-links), match rate and false match rate.

			The TTPs should enhance the linkage methods by combining deterministic linkage in the first steps using the NRN and probabilistic approaches afterwards for unlinked persons using algorithm based on other personal data. Identify subgroups of records that are more prone to linkage error and are potential sources of bias. Comparisons of linked and unlinked records can be useful to identifying where modified linkage strategies may be required for specific groups of records. Use the NRN of all individuals included in the survey, regardless of the composition of the household at one time, instead of that of the reference person first and then the other family members, in order to improve the linkage rate.
Costs	Data linkage can be expensive in terms of financial and human resources.	Government-sponsored (NIHDI) linked datasets	Make the system cost-effective by avoiding the 'linked and destroyed' philosophy and making available the linked data to other researchers under certain conditions.
Principle of proportionality respect	Means that only data that are relevant to the purpose of the study should be included to avoid re-identification of individuals.	Help from the BHIS team for the selection of BHIS variables and help from the IMA's SPOC for what concern IMA variables.	Require a deep knowledge of the data sources. Involve people with good experience of the data sources to be linked in the relevant variable selection phase. An alternative and more effective approach could be could be too ask for authorization to link both datasets completely in a first step. In a second step, each research project demands in a simplified procedure access to the relevant variables of the fully linked dataset in accordance with the proportionality principle. Such an approach is applied at Statistics Netherlands (49–51).
Infrastructures	Infrastructure needed to store and access the linked data.	The linked data was stored on the IMA server. Researchers access it through a secure remote connection using a token.	Identify where linked data can be stored securely and how it can be accessed (remote session, data extraction).

Statistical issues	Analysing linked data raises a number of statistical challenges for researchers.	Experts' advice during the statistical analysis plan, data analysis and interpretation of results.	Experts' advice useful for the statistical analysis plan, data analysis and results interpretation. Apply appropriate statistical methods of adjusting analysis for linkage bias. E.g., an extension to standard multiple imputation methods, able to handle 'partially observed' (or partially linked) data; use of population weights to account for groups or people who are more or less likely to be linked (46).
Ethical, legal and societal aspects			
Approval processes	Privacy concerns have led to policies that prevent records from being easily linked. Usually, there is a need of intuitional/ethical review boards (IRB) approval which is a long and cumbersome process.	The linkage was approved by the Information Security committee (ISC). The approval process took three and five months for the HISlink 2013 and HISlink 2018, respectively.	Consider the IRB process in the timeline for the project. Concerns about privacy led to policies that prevent records from being easily linked. Therefore, a strong case for using the data and a detailed description of how it will be protected is required when obtaining IRB approval. Since the HISlink is government-sponsored linkage project which is repeated every BHIS wave, a solution to avoid an ad hoc approval process would be to set up an "umbrella" agreement protocol for public institutions such as Sciensano, covering several years and several waves of BHIS_BCHI linkages.
Privacy and confidentiality issues: actual linkage process and principle of separation (Trusted Third Party linkage)	Once the IRB approval has been obtained, the actual linkage is itself a time-consuming process. The separation principle means a separation of the linking and analysis process. Although this principle preserves confidentiality and avoids disclosing sensitive information, it is bad for understanding the quality of linked data.	Trusted Third Party linkage, a lengthy process mainly due to the signing of an agreement between all parties involved. The whole linkage procedure took 12 months and 15 months for HISlink 2013 and HISlink 2018, respectively..	Although full separation of identifiers and attribute data has been argued to reduce the risk of re-identification, and is a valuable tool in reassuring data providers about the security of sharing their data. However, allowing linkage and analysis to take place together provides opportunities for both in-depth evaluation of linkage quality, and methodological advances in linkage technics (76,77).
Consent form	To comply with the GDPR, an effective opt-in linkage consent form have to be received.	HISlink 2013 and 2018 were not consent-based (exemptions, linkage	For planned linkage, ask for linkage consent to the survey participants, preferably at the beginning of the survey to maximise consent rate (55,68,69).

		planed before the implementation of the GDPR). However, for the next HISlink 2023, the consent of the BHIS participants was asked to link their data with existing administrative data..	For historical data linkage, certain exemptions exist. Check if the project falls under these exemptions. Assess consent bias if applicable
Outcomes			
Opportunities / limitations of linked data	The linked data is an important source for population health research and can bring enormous benefits in providing a more complete picture of the health of the population. A whole range of research possibilities exists.	Limitations of both BHIS and BCHI data remain, for instance lack of diagnostic information in the BCHI data	Include other data sources such as hospital discharge data Consider substituting HIS information by administrative data as much as appropriate (e.g., or cancer screening, reimbursed, healthcare use or reimbursed drug use).
Linkage type and sustainability	Ad hoc linkages vs. systematic linkages	Ad hoc linkage (and ad hoc approval) can threat the sustainability of the project. HISlink is based on the 'linked and destroyed philosophy' (because of a limited data retention time by researchers in the IRB approval, i.e., five years after the linkage) As a result, the return on investment in linked data may be limited.	A clear data use agreements for governmental institutions, administrations, universities allowing share and use of the linked databases for at least several years even if for perpetuity in a secure manner. Such strategies will allow to exploit the full potential of the linked data in other researches. Think about systematic linkage.
Access to the linked data		HISlink data is currently accessible to Sciensano researchers only.	Make de-identified data available to other researchers upon approval
Sample size	Small sample can prevent some analyses	Limited sample size for rare events, specific subgroup analysis	Consider subsample for specific subgroups such as low sociodemographic individuals, those with specific conditions if possible.

7.6. CONCLUSIONS

Data linkage provides important added value for public health researchers. From a public health perspective, policy makers should continue investing in data linkages; and the inclusion of other data sources such as primary care data and hospital discharge data will augment the use of the linked data to expand the evidence base for policy makers and practitioners, and can thus enrich population-based surveillance and the field of research into public health. Considering the strengths and limitations of different data sources, the opportunity to link several data sources could potentially enable a wider range of research questions to be addressed. However, linking survey data to administrative data is not without its challenges and these have to be tackled. Although some aspects of the HISlink may be specific to the Belgian context, we believe that this study has a much broader application and could be useful to researchers who plan to link health survey data with health administrative data for their respective projects.

Declarations

Ethics approval and consent to participate

This study was carried out through an individual linkage between BHIS 2013 and 2018 data and the BCHI data. The BHIS was carried out in line with Belgian privacy legislation. Both BHIS 2013 and 2018 were approved by the Ghent University Hospital ethics committee on October, 1st 2012 (opinion EC UZG 2012/658) and on December, 21 2017 (opinion EC UZG 2017/1454) respectively. Participation in the BHIS is voluntary. No formal written and signed consent was foreseen as participation was considered as consent. In addition, for the data linkage, authorization was obtained from the Belgian Information security committee (local reference: Deliberation No. 17/119 of December 19, 2017, amended on September 3, 2019 for the HISlink 2013 and local reference: Deliberation No. 20/204 of November 3, 2020 for the HISlink 2018).

Consent for publication

Not applicable.

Availability of data and materials

The survey datasets and linked health administrative data analysed in the current study are not publicly available due to restrictions in the General Data Protection Regulation (GDPR) on sensitive data such as personal health data. BHIS data contains sensitive and identifying information and must therefore only be made available upon request. Requests for data access may be made to the Social Security and Health Chamber of the Information security committee (hereinafter referred to as the "Social Security and Health Chamber"). Further information regarding the survey and the data access procedure can be found here: Health Interview Survey | Microdata request procedure | sciensano.be.

Competing interests

The authors declare that they have no competing interests.

Funding

This work did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The Belgian Health Interview Survey (BHIS) is financed by the Federal and Inter-Federated Belgian Public Health authorities. The linkage between BHIS data and the Belgian Compulsory Health Insurance data is financed by the National Institute for Health and Disability Insurance.

Authors' contributions

FB and JVdH were responsible for designing the objectives and approach of the study. FB conducted the literature searches, undertook the statistical analyses, interpreted the results, wrote the initial version of the manuscript. FB was a main contributor in writing the manuscript. JVdH, SD, RC, KDR, HVO, WVH and OB critically revised the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements

We would like to thank Statbel, the Belgian statistical office, which was responsible for BHIS sample selection and fieldwork management. Thanks to Youri Baeyens and the Inter Mutualist Agency (IMA) for their involvement in the process of data linkage.

7.7. BIBLIOGRAPHY

1. March S, Andrich S, Drepper J, Horenkamp-Sonntag D, Icks A, Ihle P, et al. Good Practice Data Linkage (GPD): A Translation of the German Version. IJERPH. 2020 Oct 27;17(21):7852.
2. Druschke D, Arnold K, Heinrich L, Reichert J, Rüdiger M, Schmitt J. Individual-Level Linkage of Primary and Secondary Data from Three Sources for Comprehensive Analyses of Low Birthweight Effects. Gesundheitswesen. 2020 Mar;82(S 02):S108–16.
3. Centre for Health Record Linkage (CHeReL). New South Wales (NSW) Gouvernment Website - Centre for Health Record Linkage. [cited 2023 Feb 9]. How record linkage works. Available from: https://www.cherel.org.au/how-record-linkage-works#:~:text=How%20record%20linkage%20works,of%20health%20events%20for%20individuals.
4. Brook EL, Rosman DL, Holman CDJ. Public good through data linkage: measuring research outputs from the Western Australian Data Linkage System. Australian and New Zealand Journal of Public Health. 2008 Feb;32(1):19–23.
5. Tew M, Dalziel KM, Petrie DJ, Clarke PM. Growth of linked hospital data use in Australia: a systematic review. Aust Health Review. 2017;41(4):394.
6. Young A, Flack F. Recent trends in the use of linked data in Australia. Aust Health Review. 2018;42(5):584.
7. Maret-Ouda J, Tao W, Wahlin K, Lagergren J. Nordic registry-based cohort studies: Possibilities and pitfalls when combining Nordic registry data. Scand J Public Health. 2017 Jul;45(17_suppl):14–9.
8. Haneef R, Delnord M, Vernay M, Bauchet E, Gaidelyte R, Van Oyen H, et al. Innovative use of data sources: a cross-sectional study of data linkage and artificial intelligence practices across European countries. Arch Public Health. 2020 Dec;78(1):55.
9. March S. Individual Data Linkage of Survey Data with Claims Data in Germany—An Overview Based on a Cohort Study. IJERPH. 2017 Dec 9;14(12):1543.
10. Hall HI, Van Den Eeden SK, Tolsma DD, Rardin K, Thompson T, Hughes Sinclair A, et al. Testing for prostate and colorectal cancer: comparison of self-report and medical record audit. Preventive Medicine. 2004 Jul;39(1):27–35.
11. Van der Heyden J, Charafeddine R, De Bacquer D, Tafforeau J, Van Herck K. Regional differences in the validity of self-reported use of health care in Belgium: selection versus reporting bias. BMC Med Res Methodol. 2016 Dec;16(1):98.
12. Van der Heyden J, Van Oyen H, Berger N, De Bacquer D, Van Herck K. Activity limitations predict health care expenditures in the general population in Belgium. BMC Public Health. 2015 Dec;15(1):267.

13. Mimilidis Hélène, Demarest Stefaan, Tafforeau Jean, Van der Heyden Johan. Projet de couplage de données issues de l'Enquête de Santé 2008 et des Organismes Assureurs. Bruxelles, Belgique; 2014 Mai. Report No.: D/2014/2505/32.
14. Holman CDJ, Bass AJ, Rouse IL, Hobbs MST. Population-based linkage of health records in Western Australia: development of a health services research linked database. Australian and New Zealand Journal of Public Health. 1999 Oct;23(5):453–9.
15. Holman CDJ, Bass JA, Rosman DL, Smith MB, Semmens JB, Glasson EJ, et al. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. Aust Health Review. 2008;32(4):766.
16. Mirel LB. The NCHS Data Linkage Program: Leveraging the nation's health data for evidence-based decision making. In 2020 [cited 2022 Jun 27]. p. 28. Available from: https://www.cdc.gov/nchs/data/datalinkage/Data-Linkage-Webinar.pdf
17. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. Big Data & Society. 2017 Dec;4(2):205395171774567.
18. Harron K. Data linkage in medical research. bmjmed. 2022 Mar;1(1):e000087.
19. Harron K, Gilbert R, Cromwell D, van der Meulen J. Linking Data for Mothers and Babies in De-Identified Electronic Health Data. Gebhardt S, editor. PLoS ONE. 2016 Oct 20;11(10):e0164667.
20. Demarest S, Van der Heyden J, Charafeddine R, Drieskens S, Gisle L, Tafforeau J. Methodological basics and evolution of the Belgian health interview survey 1997–2008. Arch Public Health. 2013 Dec;71(1):24.
21. Van der Heyden J. Validity of the Assessment of Population Health and Use of Health Care in a National Health Interview Survey [Internet]. [Ghent, Belgium]: Ghent University - Faculty of medicine and health sciences; 2017 [cited 2023 Feb 9]. Available from: https://biblio.ugent.be/publication/8523878
22. Berete F, Van der Heyden J, Demarest S, Charafeddine R, Tafforeau J, Van Oyen H, et al. Validity of self-reported mammography uptake in the Belgian health interview survey: selection and reporting bias. European Journal of Public Health. 2021 Feb 1;31(1):214–20.
23. KORA Study Group, Hunger M, Schwarzkopf L, Heier M, Peters A, Holle R. Official statistics and claims data records indicate non-response and recall bias within survey-based estimates of health care utilization in the older population. BMC Health Serv Res. 2013 Dec;13(1):1.
24. Devos C, Cordon A, Lefevre M, Obyn C, Renard F, Bouckaert N, et al. Performance of the Belgian health system–report 2019. Health Services Research (HSR). Brussels: Belgian Health Care Knowledge Centre (KCE); 2019.

25. Noordhout CMD, Devos C, Adriaenssens J, Bouckaert N, Ricour C, Gerkens S. Health system performance assessment: care for people living with chronic conditions.
26. Bouckaert N, Maertens de Noordhout C, Van de Voorde C. Health System Performance Assessment: how equitable is the Belgian health system? [Internet]. Brussels: Belgian: Health Services Research (HSR). Health Care Knowledge Centre (KCE); 2020 [cited 2022 Jun 27] p. 105. Report No.: KCE Reports 334. D/2020/10.273/30. Available from: https://kce.fgov.be/sites/default/files/2021-11/KCE_334_Equity_Belgian_health_system_Report.pdf
27. Berete F, Demarest S, Charafeddine R, Bruyère O, Van der Heyden J. Comparing health insurance data and health interview survey data for ascertaining chronic disease prevalence in Belgium. Arch Public Health. 2020 Dec;78(1):120.
28. Maetens A, De Schreye R, Faes K, Houttekier D, Deliens L, Gielen B, et al. Using linked administrative and disease-specific databases to study end-of-life care on a population level. BMC Palliat Care. 2016 Dec;15(1):86.
29. Berete F, Demarest S, Charafeddine R, Ridder K, Vanoverloop J, Oyen H, et al. Predictors of Nursing Home Admission in the Older Population in Belgium [Internet]. In Review; 2022 Jan [cited 2022 Mar 3]. Available from: https://www.researchsquare.com/article/rs-1169480/v1
30. Van der Heyden J, Berete F, Renard F, Vanoverloop J, Devleesschauwer B, De Ridder K, et al. Assessing polypharmacy in the older population: Comparison of a self-reported and prescription based method. Pharmacoepidemiol Drug Saf. 2021 Dec;30(12):1716–26.
31. Agence InterMutualiste -InterMutualistisch Agentschap (AIM-IMA). Agence InterMutualiste -InterMutualistisch Agentschap [Internet]. [cited 2021 Jul 26]. Available from: https://www.ima-aim.be/-Donnees-de-sante-
32. Finaba Berete, Johan Vander Heyden, Stefaan Demarest, Rana Charafeddine. Couplage des données de l'enquête de santé avec les données des organismes assureurs - Hislink 2013 Méthodologie et étude comparative sur la prévalence des maladies chroniques. 2020 Apr.
33. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [Internet]. 2016 [cited 2023 Sep 28]. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&qid=1695900759326
34. GDPR-in-short. European legislation related to data standards [Internet]. 2021 [cited 2023 Oct 24]. Available from: https://www.polisnetwork.eu/wp-content/uploads/2021/03/GDPR-in-short2.pdf

35. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. International Journal of Epidemiology. 2017 Oct 1;46(5):1699–710.
36. Williams N, Hermans K, Stevens T, Hirdes JP, Declercq A, Cohen J, et al. Prognosis does not change the landscape: palliative home care clients experience high rates of pain and nausea, regardless of prognosis. BMC Palliat Care. 2021 Dec;20(1):165.
37. Austin PC. Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research. Communications in Statistics - Simulation and Computation. 2009 May 14;38(6):1228–34.
38. Saunders NR, Janus M, Porter J, Lu H, Gaskin A, Kalappa G, et al. Use of administrative record linkage to measure medical and social risk factors for early developmental vulnerability in Ontario, Canada. IJPDS [Internet]. 2021 Feb 11 [cited 2022 Mar 3];6(1). Available from: https://ijpds.org/article/view/1407
39. Berete F, Van der Heyden J, Demarest S, Van Oyen H, Charafeddine R, Bruyère O. Effectiveness of protective measures on dental care utilization: analysis from linked database. . European Journal of Public Health. 2020 Sep 30;30(5).
40. Berete F, Charafeddine R, Demarest S, Heyden JV der, Gisle L, Van Den Broucke S, et al. Does health literacy mediate the relationship between socioeconomic status and health(-related) outcomes in the Belgian adult population? Will be submitted to BMC Public Health. 2023;
41. Gorasso V, Moyersoen I, Van der Heyden J, De Ridder K, Vandevijvere S, Vansteelandt S, et al. Health care costs and lost productivity costs related to excess weight in Belgium. BMC Public Health. 2022 Sep 6;22(1):1693.
42. Van der Heyden J, Berete F, Devleesschauwer B, De Ridder K, Bruyère O, Renard F, et al. Association between polypharmacy and mortality in the community-dwelling older population: a data linkage study. International Journal of Epidemiology. 2021 Sep;50:239–239.
43. Harron K, Doidge J. Challenges and opportunities in using administrative data linkage for research: the importance of quality assessment for understanding bias [Internet]. 2020 Jan; UCL Great Ormond Street Institute of Child Health. Available from: https://www.ucl.ac.uk/population-health-sciences/sites/population_health_sciences/files/1-nash-mina_katieharron_jan2020.pdf
44. Ludvigsson JF, Otterblad-Olausson P, Pettersson BU, Ekbom A. The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research. Eur J Epidemiol. 2009 Nov;24(11):659–67.
45. Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR. Linking data for health services research: a framework and instructional guide. 2014;
46. Harron K, Mackay E, Elliot M. An introduction to data linkage. 2016;

47. Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang LC, Smith P, et al. GUILD: GUidance for Information about Linking Data sets†. Journal of Public Health. 2018 Mar 1;40(1):191–8.

48. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, et al. Data Linkage: A powerful research tool with potential problems. BMC Health Serv Res. 2010 Dec;10(1):346.

49. Sediq R, Van Der Schans J, Dotinga A, Alingh RA, Wilffert B, Bos JH, et al. Concordance assessment of self-reported medication use in the Netherlands three-generation Lifelines Cohort study with the pharmacy database iaDB. nl: The PharmLines initiative. Clinical epidemiology. 2018;981–9.

50. van Brug HE, Rosendaal FR, van Steenbergen LN, Nelissen RG, Gademan MG. Data linkage of two national databases: Lessons learned from linking the Dutch Arthroplasty Register with the Dutch Foundation for Pharmaceutical Statistics. Plos one. 2023;18(3):e0282519.

51. Applying for linked data from the Lifelines Cohort Study and IADB.nl database [Internet]. 2021. Available from: http://wiki-lifelines.web.rug.nl/lib/exe/fetch.php?media=pharmlines_procedures_20210415.pdf

52. Jutte DP, Roos LL, Brownell MD. Administrative Record Linkage as a Tool for Public Health Research. Annu Rev Public Health. 2011 Apr 21;32(1):91–108.

53. European Parliament and European Council. Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) [Internet]. 2022 [cited 2022 Sep 27]. Available from: https://data.consilium.europa.eu/doc/document/PE-85-2021-INIT/en/pdf

54. Sakshaug JW, Couper MP, Ofstedal MB, Weir DR. Linking Survey and Administrative Records: Mechanisms of Consent. Sociological Methods & Research. 2012 Nov;41(4):535–69.

55. Sakshaug JW, Schmucker A, Kreuter F, Couper MP, Holtmann L. Respondent understanding of data linkage consent. Survey Methods: Insights from the Field (SMIF). 2021;

56. van Veen EB. Observational health research in Europe: understanding the General Data Protection Regulation and underlying debate. European Journal of Cancer. 2018 Nov;104:70–80.

57. Hafferty JD, Campbell AI, Navrady LB, Adams MJ, MacIntyre D, Lawrie SM, et al. Self-reported medication use validated through record linkage to national prescribing data. Journal of Clinical Epidemiology. 2018 Feb;94:132–42.

58. Richardson K, Kenny RA, Peklar J, Bennett K. Agreement between patient interview data on prescription medication use and pharmacy records in those aged older than 50 years varied by therapeutic group and reporting of indicated health conditions. Journal of Clinical Epidemiology. 2013 Nov;66(11):1308–16.

59. Rosella LC, Manuel DG, Burchill C, Stukel TA, for the PHIAT-DM team. A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT). Journal of Epidemiology & Community Health. 2011 Jul 1;65(7):613–20.

60. Rosella LC, Fitzpatrick T, Wodchis WP, Calzavara A, Manson H, Goel V. High-cost health care users in Ontario, Canada: demographic, socio-economic, and health status characteristics. BMC Health Serv Res. 2014 Dec;14(1):532.

61. Gorman E, Leyland AH, McCartney G, White IR, Katikireddi SV, Rutherford L, et al. Assessing the Representativeness of Population-Sampled Health Surveys Through Linkage to Administrative Data on Alcohol-Related Outcomes. American Journal of Epidemiology. 2014 Nov 1;180(9):941–8.

62. Meyer BD, Mittag N. Combining administrative and survey data to improve income measurement. Administrative Records for Survey Methodology. 2021;297–322.

63. Morgan K, Page N, Brown R, Long S, Hewitt G, Del Pozo-Banos M, et al. Sources of potential bias when combining routine data linkage and a national survey of secondary school-aged children: a record linkage study. BMC Med Res Methodol. 2020 Dec;20(1):178.

64. Linnenkamp U, Gontscharuk V, Brüne M, Chernyak N, Kvitkina T, Arend W, et al. Using statutory health insurance data to evaluate non-response in a cross-sectional study on depression among patients with diabetes in Germany. International Journal of Epidemiology. 2020 Apr 1;49(2):629–37.

65. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future: Health Services Research and Data Linkages. Health Services Research. 2010 Oct;45(5p2):1468–88.

66. Maddocks J, Mathieu L, Richards R, Saelaert M, Van Hoof W. tehdas-healthy-data-an-online-citizen-consultation-about-health-data-reuse-intermediate-report.pdf [Internet]. 2022 Jun [cited 2022 Sep 27]. Available from: https://tehdas.eu/app/uploads/2022/07/tehdas-healthy-data-an-online-citizen-consultation-about-health-data-reuse-intermediate-report.pdf

67. Marie Thornby LC. Collecting Multiple Data Linkage Consents in a Mixed-mode Survey: Evidence from a large-scale longitudinal study in the UK. 2018 [cited 2022 Sep 27]; Available from: https://surveyinsights.org/?p=9734

68. Sakshaug JW, Vicari BJ. Obtaining Record Linkage Consent from Establishments: The Impact of Question Placement on Consent Rates and Bias. Journal of Survey Statistics and Methodology. 2018 Mar 1;6(1):46–71.

69. Sakshaug JW, Schmucker A, Kreuter F, Couper MP, Singer E. The Effect of Framing and Placement on Linkage Consent. Public Opinion Quarterly. 2019 Jul 19;83(S1):289–308.

70. European Commission. REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space [Internet]. 2022 [cited 2023 Feb 9]. Available from: https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF
71. McLennan S, Celi LA, Buyx A. COVID-19: Putting the General Data Protection Regulation to the Test. JMIR Public Health Surveill. 2020 May 29;6(2):e19279.
72. Kiseleva A, De Hert P. Creating a European Health Data Space: Obstacles in Four Key Legal Area. EPLR. 2021;5:21.
73. Dibben C, Elliot M, Gowans H, Lightfoot D, Data Linkage Centres. The data linkage environment. In: Methodological Developments in Data Linkage Chapter 3. Harron K, Dibben C and Goldstein H (eds). London: Wiley; 2015.
74. Jones KH, Ford DV, Jones C, Dsilva R, Thompson S, Brooks CJ, et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation. Journal of Biomedical Informatics. 2014 Aug;50:196–204.
75. Boyd JH, Ferrante AM, O'Keefe CM, Bass AJ, Randall SM, Semmens JB. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. BMC Health Serv Res. 2012 Dec;12(1):480.
76. Aldridge RW, Shaji K, Hayward AC, Abubakar I. Accuracy of Probabilistic Linkage Using the Enhanced Matching System for Public Health and Epidemiological Studies. Pacheco AG, editor. PLoS ONE. 2015 Aug 24;10(8):e0136179.
77. Harron K, Goldstein H, Wade A, Muller-Pebody B, Parslow R, Gilbert R. Linkage, Evaluation and Analysis of National Electronic Healthcare Data: Application to Providing Enhanced Blood-Stream Infection Surveillance in Paediatric Intensive Care. Trotter CL, editor. PLoS ONE. 2013 Dec 20;8(12):e85278.

CHAPTER 8. GENERAL DISCUSSION AND RECOMMENDATIONS

8.1. INTRODUCTION

Population-based surveys such as the BHIS are essential tools to provide information on population health. However, the validity of self-reported information through surveys is a concern due to the associated selection and reporting bias. Data linkage can play a crucial role in obtaining further insights about the validity of self-reported information.

In addition to these validity issues (as a result of selection and reporting bias), population-based surveys are also facing other challenges due to the increasing need for more comprehensive data to answer complex research questions. However, increasing the number of questions in population surveys would result in a high workload for interviewers and a significant burden on respondents in the case of face-to-face data collection due to the length of questionnaires. This also leads to dropouts resulting in missing data and lower response rates, which both affect data quality. Data linkage is a useful and efficient approach for obtaining more complete data without increasing the length of the questionnaires.

The BHIS and BCHI data are important population-based data sources in Belgium. Linking the two data sources (HISlink) allows on the one hand a validation of some of the survey data, and results on the other hand in a richer database which offers new research opportunities useful to public health authorities.

Based on the use case of the HISlink, the overarching aim of this thesis was to investigate the potential benefits and opportunities of linking health survey data with health insurance data for public health research. More specifically, the following areas were covered:

1) To explore the fundamental concepts of data linkage, a literature review was undertaken to cover the following questions: What is data linkage? What are commonly the types of linked data? What methods have been used to link data? What are the challenges and the legal issues? How to assess the quality of linked data?

Then, the following two research questions were examined:

2) To what extent can linked data be used to assess data validity?

This question was explored for three topics: self-reported mammography uptake, chronic diseases and polypharmacy.

3) To what extent can linked data be used to respond to policy-relevant questions which cannot be addressed with each of the sources separately?

This question was explored for two policy-relevant research questions: what are predictors of nursing home admission among the older population? What is the mediating effect of health literacy in the relationship between socioeconomic status and health outcomes?

This chapter continues with a brief recapitulation of the main findings of the thesis. Next it moves on to the strengths and limitations. The chapter continues with future perspectives, implications and recommendations, and ends with a final conclusion.

8.2. SUMMARY OF THE MAIN FINDINGS

8.2.1. Summary of the results of the literature review

The main results of the literature review are summarised in Table 8.1.

Table 8.1: Summary of the literature on data linkage

Questions	Main findings
What is data linkage?	• Data linkage brings together information that relates to the same individual, family, place or event from different data sources.
What are commonly the types of linked data?	• Varying data sources within the context of health research can be linked, including survey data (both cross-sectional and longitudinal) and administrative data. • Data from health interview surveys, health examination surveys and social surveys are the most commonly used as initial data source for performing a data linkage. • Common sources of health-related administrative data involved in data linkages are health insurance claims data, hospital discharge data, prescription drugs data, medical records, disease-specific registries.
What are the linkage methods?	• The content and quality of the data sources to be linked play an important role in the choice of linkage methods. • Deterministic methods are simplest and best suited to 'perfect' data where there are unique personal identifiers or highly discriminating linkage keys. • Probabilistic methods are more complex and can be adapted to imperfect data.
What are the challenges and the legal issues?	• Privacy and confidentiality issues remain the key concerns.
How to assess the quality of linked data?	• Linkage errors pose the greatest threat to the quality of the linked data and ultimately lead to information bias and selection bias. • Care must be taken to assess the quality of the linkage in order to provide reliable results. • Several methods are proposed to assess the quality of the linked data including standard metrics (e.g. match rate, recall, precision, etc.) or more elaborated approaches (e.g. comparison with gold standard, sensitivity analysis, comparison linked vs. unlinked data, quality control check, etc.). • Researchers should validate the linked data before undertaking any analysis using them.

8.2.2. Summary of findings from studies carried out to address the research questions

8.2.2.1. How valid is the information from the BHIS source as compared with the information from the BCHI source?

The validity of BHIS information was addressed in chapter 4. The first article in the chapter on mammography uptake examined the criterion validity of BHIS 2013 information on this topic, using BCHI data as the gold standard. The other two articles, on ascertaining the prevalence of a selection of CDs and on polypharmacy, compared BHIS and BCHI data and assessed their complementarity. Table 8.2 summarises the main findings of the case studies.

Table 8.2: Overview of the main findings regarding validity studies

Case study 1: mammography uptake
• By relying on survey data there is a significant overestimation of participation in mammography screening within the target group • Both selection and reporting bias have an impact on the validity of BHIS data regarding this topic • There is a substantial difference in validity of mammography uptake across population subgroups
Case study 2: CD's comparison
• Estimating the prevalence of chronic diseases from data on reimbursed drugs alone has significant limitations. • Caution should be exercised when using indicators based on these data alone to estimate the prevalence of chronic diseases • There is good agreement between survey- and health insurance-based estimates for some CDs such as diabetes, Parkinson's disease and thyroid disorders, but a poor agreement for COPD and asthma.
Case study 3: Polypharmacy
• The BHIS data can be used to estimate polypharmacy if appropriate survey instruments are used. • BHIS data are better suited to this purpose than the BCHI data. • There is no difference in the determinants of moderate polypharmacy according to the source of the outcome.

On the whole, the data collected as part of an HIS differs from that found in administrative data sources, and this varies according to the specific topics under consideration and the characteristics of the survey participants. Although both data sources have their advantages, relying solely on one would result in an estimate of the indicator in question that may be less comprehensive. Therefore, objective data should be combined with survey data wherever possible.

8.2.2.2. To what extent can linked data be used to better respond to policy-relevant questions? Results from 2 case studies

Research topics related to this question were addressed in chapters 5 and 6. Chapter 5 showed how the linkage of BHIS data with longitudinal BCHI data can be used to estimate the cumulative risk of NHA among the older population of 65+ years and its predictors in Belgium. While chapter 6 presented a case study investigating the mediating effects of HL on the relationship between education, income and a selected health and health-related outcomes (HRO). Table 8.3 summaries the main findings of these studies

Table 8.3: Overview of the main findings regarding studies on policy-relevant questions

Case study 1: Nursing home admission (NHA)
• The cumulative risk of NHA was 1.4%, 5.7% and 13.1% at, respectively 1 year, 3 years and 5 years of follow-up • The factors predicting NHA are multifactorial: a higher age, living situation (social supports), history of falls, urinary incontinence, physical chronic conditions and mental disorders such as Alzheimer's disease, appeared as strong predictors of NHA. • Preventing falls, managing urinary incontinence at home and providing appropriate and timely management of limitations, depression and Alzheimer's disease would delay the onset of NHA.
Case study 2: Mediation effects of health literacy (HL)
• HL acts as mediator in the relationship between education, income and health related outcomes in a range of domains: preventive care, health status, health behaviour, use of medicines. • The effect of HL is rather limited and varied across the health related outcomes. • The mediating effect of HL accounted significantly for 2% to 15.4% of the total effect, suggesting that improving HL might reduce SES disparities in these areas

The results from these two case studies showed that in public health, to answer certain research questions the use of multiple data sources is required. In such cases,

data linkage is a powerful tool for obtaining a richer database from which to carry out the necessary analyses. For example, our studies showed that linking survey data with health administrative data has enabled the conducting of research that would otherwise have been impossible. Specifically, while some information can only be extracted from administrative data sources (for example, date of entry into the nursing home), other information can only be obtained through health surveys (such as health status, health behaviour, social support). Furthermore, thanks to the linked data, the researcher can also choose to combine information from the two sources in order to obtain a more accurate indicator, or to choose the source of the information according to the confidence placed in the source of this information.

8.2.3. Lessons learned with respect to the actual linkage

The lessons learned from this thesis go beyond the results of the studies carried out. The main lessons learned outside of the studies' findings are summarised below.

- **Ethical, legal and societal aspects**

The most important challenges were the privacy and confidentiality issues. Because of these issues, an institutional review board (IRB) approval was required to link the data. However, such IRB approval processes were complex and time-consuming, For both HISlink 2013 and HISlink 2018, it took several months to get the IRB approval. Moreover, to preserve privacy and prevent the disclosure of sensitive information, the linkage was carried out by a complex process involving two TTPs. This approach led to a significant delay in the linkage process. Some administrative steps such as simply obtaining the signature of an official took more time than anticipated. With the implementation of the GDPR in 2018, new decision-making bodies were established for the authorisation of data linkage, and privacy and confidentiality issues were redefined.

- **Technical and operational issues of the linkage**

The availability of a unique personal identifier (the National Register Number) has greatly facilitated the linkage process. Although the NRN guided the choice of the simplest linkage method (i.e. deterministic linkage), this does not rule out the risk of linkage errors.

Next, the principle of proportionality implied a careful selection beforehand of all the data that will be required, which supposes in-depth knowledge of the two data sources concerned. The more information the two data sources contain, the more difficult this task becomes, making the approach sub-optimal. An alternative, and perhaps better approach, might be to seek permission to link the two datasets completely in the first instance. In the second stage, each research project applies, under a simplified procedure, for access to the relevant variables in the fully linked dataset, in accordance with the principle of proportionality. This is essentially what is done at Statistics Netherlands (1–3).

The linkage was not always possible for everyone. A possible explanation for this would be that only the NRN of the reference person was available and the others had to be found on the basis of household composition and socio-demographic characteristics. This approach is probably linked to the BHIS sampling strategy. However, BHIS household composition may differ from BCHI household composition or may change over time. Therefore, including the NRN of all individuals included in the survey, regardless of household composition would probably improve the linkage.

The BCHI data is only available two years after consumption, meaning that the linkage can only be made with a two-year delay which precludes 'real time' linkage. Data availability should be accelerated in the short to medium term given the widespread use of electronic billing.

Furthermore, with the limited sample size of the BHIS (about 10,000 participants), subgroup analysis is impossible or yields inaccurate results, for example for rare events or specific subgroups.

The use of administrative data for epidemiological purposes was challenging as they are not meant for this. In addition, there was a need for analysts with expertise in the two types of data sources because of differences, for example, in the definition of cases in the two data sources.

- **The linkage of both data sources combines their strengths but does not overcome all the weaknesses**

The HISlink use case highlighted the fact that linking survey data and health insurance data combines their strengths while compensating for certain weaknesses. However,

this does not eliminate all biases or resolve all weaknesses. For example, reporting bias in survey data persists in linked data.

- **Need for good collaboration between all partners involved**

The linkage process was complex, involving several partners whose roles were clearly defined at each stage of the process. Good collaboration between the partners was therefore essential to the success of the project.

8.3. STRENGTHS OF THE THESIS

- **Data component**

The main strength of this thesis lies in the fact that is based on data from a representative sample of the population obtained through a systematic linkage between two important and complementary population-based data sources in Belgium, namely the BHIS and the BCHI. The use of BCHI has been particularly useful because of the compulsory nature of this insurance, which covers comprehensive information on healthcare use for almost the entire population.

Through the implementation of the HISlink, useful experience was obtained for future linkages. Although based on Belgian data, the HISlink project should be seen as an example that also provides useful information for future linkages in another context. Indeed, HISlink covers survey and administrative data sources, which are among the common sources of data evolved in data linkage. Therefore, our recommendations could be useful for future linkages in EU countries where it is possible to link HIS data with health insurance data covering comprehensive information on the use of health care for the (whole) population.

- **Concomitant assessment of selection and reporting bias**

A particular strength of the study on the validity of self-reported mammography uptake was that we assessed concomitantly the selection (through a comparison with EPS, a random sample of BCHI data) and the reporting bias, which made it possible to estimate the relative importance of both biases. This is particularly of importance because one can use the bias information to introduce a correction factor for BHIS data.

- **Rigorous treatment of missing data**

This study was also strengthened by the correct treatment of item non-response. In fact, in several studies included in this thesis, item non-response was handled by multiple imputation. This reduced item-nonresponse bias and improved the generalisability of our results.

8.4. LIMITATIONS OF THE THESIS

- **Assessment of limited topics**

The topic of this thesis is quite broad, offering a wide range of research possibilities. Therefore, certain choices had to be made. As a result, only a limited number of themes have been addressed. Our choices were guided by the work already done, the relevance of the topics, and added value for public health (relevance for the commissioner of the linkage, relevance with respect to societal challenges (ageing population)), but also the feasibility in the context of the data available in both databases. As a consequence, the conclusions drawn are restricted to the findings of those topics and the implications for public health are based solely on the results of the topics studied.

- **Extrapolation to other EU countries**

The results of this study are based on Belgian data, which suggests that the conclusions are more related to the Belgian context. The Belgian healthcare system structure may differ from that of other EU countries. In addition, the contents of health insurance data could be different from one country to another. Therefore, care should be taken when extrapolating the results.

- **Data linkage to two specific data sources**

Another limitation is that we focused only on the linkage with the BCHI and not on other sources of health-related administrative data. This has, to some extent, led to a restriction in the choice of studies, indicators or methods used in this thesis. It would therefore be interesting to consider extending the linkage to other sources of administrative data such as hospital discharge data, social security data, environmental data, etc.

- **Other limitations related to specific studies**

Limitations related to each of the specific studies (e.g. lack of diagnostic codes) are discussed in the corresponding chapters.

8.5. FUTURE PERSPECTIVES

The results of this thesis unambiguously demonstrate the benefits of linking HIS and health insurance data. However, the results of this thesis could serve as a basis for further research that expands the possibilities offered by a linkage between HIS and administrative data in general. Future prospects can be divided into methodological research that goes beyond issues of validity; specific research topics that can be addressed when extending the linked data to include other administrative data such as mortality data, hospital discharge data, etc.; as well as contextual developments in the secondary use of data, including routine linkage of administrative data and the use of real-word data (RWD).

8.5.1. Methodological research

8.5.1.1. Assessment of the representativeness of HIS data

One of the main advantages of HIS is that it allows trends to be monitored and is often carried out on a large sample presumed to be representative of the general population. Although post-stratification weights are applied to ensure that the results are representative of the population, it is not certain that representativeness is guaranteed. As our study on breast cancer screening and previous studies on healthcare use have shown (4), BHIS information suffers from both selection bias and reporting bias. Selection bias can affect the representativeness of the data and may have an impact on trends. Moreover, in view of the declining participation rates in surveys in many countries, the representativeness of survey-based results remains an important issue. A study in Finland showed that increasing non-participation over time can affect smoking trends (5). This illustrates the need to assess the extent to which the HIS data is indeed representative of the population from which the sample is drawn and the extent to which this representativeness evolves over time. Linked HIS and administrative data offers a significant increase in the number of auxiliary variables that may be used to assess or adjust for non-response bias in survey data and therefore improve the representativeness of the results (6). In addition, as HIS

samples are generally drawn from an official sampling frame (e.g. the national register), a linkage to the population as a whole allows the study sample to be compared with the population from which it is drawn.

8.5.1.2. Estimation of the correction factor for self-reported information

The individual linkage of survey data with health insurance data over several survey waves makes it possible to calculate correction factors (objective administrative data/subjective self-reported data) which could be useful for bias adjustment (7). The individual self-reported data are then multiplied with the corresponding correction factors to obtain more reliable estimates for future HISs. As reporting bias may change over time, correction factors should therefore be updated regularly (8).

8.5.1.3. Evaluation of the impact of linkage errors

The assessment of the quality of the linked data in Chapters 3 and 7 showed that the linkage rate is unevenly distributed between population subgroups, which may distort the results. Because of the use of a unique personal identifier, i.e. the national register number, we may assume that the linkage errors are related to missed matches (records from the same individual fail to link) rather than to false matches (records from different individuals link erroneously). Although the effect of such linkage errors has been anticipated and the results of the analysis are adjusted for population characteristics that are related to the linkage error (e.g. age, income, education, etc.) to produce estimates that are closer to the true value (9), it is important to assess the impact of these errors on the results of studies based on the linked data. Sensitivity analysis (9,10) or quantitative bias analysis can be used to evaluate how the linkage errors affect the results and how to adjust for them (9,11,12).

Other examples of the use of linked HIS and administrative data to solve methodological problems are presented in chapter 1 (section 1.3.4).

8.5.2. Further research topics that can be addressed when linking HIS to administrative data

8.5.2.1. Continue with work already carried out in order to confirm or update findings

Our work on CDs indicators in the BCHI (pseudopathologies) has highlighted certain limitations to the use of these indicators for ascertaining cases of CDs in the general population. Because of their limitations (uncertainty about the difference between pseudopathology and disease, lack of updating of obsolete definitions, hospital drugs not taken into account when determining pseudopathologies), these indicators have been replaced by Pharmacy Cost Groups (PCGs) based on the Dutch Pharmacy-based Cost Group model managed by the National Institute for Health Care, which provides an annual update. The PCG model uses specific types of drugs prescribed to individuals in a reference year as markers of CD, which are then used to adjust capitation payments to their sickness fund in the subsequent year (13). The linkage with other health administrative data, such as primary care data (which includes diagnoses), can be used to refine algorithms that assess people with specific CDs in a health insurance database. This will be important in a wider context, as in many countries, such as Belgium, health insurance data do not include diagnostic data.

8.5.2.2. The potential of linkages to reduce socio-economic differences in healthcare consumption and drug use

Certain domains that were not covered in this thesis, such as socio-economic determinants, can be explored. When possible, there is a need to invest in linkages with databases that enable the study of socioeconomic inequalities. For instance, it is obvious that fiscal data yield better data on income than self-reported data. So, a linkage with administrative data such as labour market and social security data can enable e.g. more accurate investigation of socioeconomic inequalities in health using income as proxy of SES.

8.5.2.3. The potential of linkages to conduct research with composite indicators

The linkage between the HIS and administrative data can be used to construct combined or composite indicators that are more accurate than information from separate data sources. Taking the example of our NHA study, several individuals' background characteristics (e.g. Alzheimer's disease, urinary incontinence) were constructed using both HIS information and administrative data. It is important to have such a composite indicator because a certain number of people suffering from Alzheimer's disease may not be taking the specific drugs, just as some may be taking the specific drugs but do not report to be suffering from the disease during the survey. Similarly, regarding urinary incontinence, a TILDA study showed that of the participants who reported incontinence during the survey, only 3 out of 5 had informed a healthcare provider (14). Therefore, relying on information from a single database would result in a poor estimate of the prevalence of this condition and have serious implications for any analysis derived from these data (15). If we take the example of polypharmacy, the HIS data includes non-reimbursed medicines, while the administrative data includes medicines that survey participants may not have shown to the interviewers. Consequently, a combined indicator of the two data sources would provide a more reliable picture of actual drug consumption.

8.5.2.4. Specific research opportunities

As indicated above, it is clear that there are still a number of research opportunities that can be carried out using the link between HIS and health insurance data. However, some studies require more comprehensive data than the current linked data. This thesis can be seen as a first step towards more comprehensive data linkages. To ensure that the benefits of data linkage are fully maximised, it is important to consider the inclusion of other administrative data such as hospital discharge data, mortality data, environmental data, the primary electronic medical record (EMR), etc. For example, extending linked data to hospital discharge data could help target internal quality improvement efforts for specific patient groups (e.g. preventive care for diabetics) or help assess the determinants of hospitalisation and understand the underlying factors that influence length of hospitalisation. A linkage with the EMR may also be useful for studying appropriate polypharmacy, for example. However, in some countries such as Belgium, there is currently no integrated primary EMR. Only a few

sentinel networks exist, such as the Intego database. For the future, consideration needs to be given to establishing a legal framework for such an integrated database. Other research possibilities are presented in chapter 1 (section 1.3.5).

8.5.3. Routine linkage of administrative data

With the digitalisation of health-related administrative data and the increased use of artificial intelligence, the linkage of routine administrative data is increasingly used in epidemiological research (16,17). Linking multiple sources of administrative data could be a valuable source of information for policymakers. In addition, given the challenge of collecting traditional data such as survey data, routine linkage of administrative data will increase further. Although administrative data cannot replace survey data for some information, it has the advantage of being readily available and continually updated.

For research purposes, administrative data have the advantage of offering detailed and accurate information, complete coverage of the populations of interest (allowing detailed analyses of sub-groups), data on the same individuals over long periods and low cost compared with survey data. Therefore, routine linkage of administrative data could contribute to the development of population cohort databases and health platforms that could be useful for answering specific research questions. For example, using a US-based electronic medical record dataset linked to claims, Ward et al. (2022) examined the relative risk of thromboembolic events resulting from COVID-19 as compared to influenzae in a large retrospective cohort (18).

Another example concerns birth cohorts in epidemiological studies. Birth cohorts are more appropriate for studying the causal relationship between potential risk factors in the prenatal or postnatal period and the health status of the newborn through to childhood and, ultimately, adulthood (19). However, recruiting and actively monitoring mothers and children is time-consuming and requires significant resources. In addition, the sample size of such a cohort is generally limited and there is a risk of subjects being lost to follow-up, which reduces statistical power and may lead to selection bias. This is why linkage of the medical birth register with the administrative medical records of mothers and babies is increasingly used in countries with universal healthcare systems, enabling researchers to identify large, unselected populations at

birth and to reconstruct the relevant characteristics and care pathways of mothers and newborns (19).

Larcin et al (2023) also linked data from national dispensing data with birth and death certificates and data on hospital stays over a period of 7 years, in order to explore the medication exposure during pregnancy (20).

8.5.4. Real-word data: a ‘new’ transition

The increased use of the internet, social media, wearable devices, e-health services, and other technology-driven services in medicine and healthcare has led to the rapid generation of various types of digital data, providing a valuable data source beyond the confines of traditional clinical trials, epidemiological studies, and lab-based experiments (21).

Real-word data (RWD) in the medical and healthcare field “are the data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources (22). These sources range from data derived from electronic health records, medical claims data, data from product or disease registries, and data gathered from other sources (such as digital health technologies: mobile devices, wearables such a pedometers and smart watches) that can inform on health status (22) to social media platforms. RWD is an emerging area of interest across the healthcare spectrum (23) and its progress is closely related to digitization, especially of medical administrative data and medical records (24).

Epidemiology and pharmacoepidemiology frequently use RWD from healthcare teams to inform research (25). RWD are generating greater interest in recent times despite not being new. There are various purposes of the RWD analytics in medical research as follows: effectiveness and safety of medical treatment, epidemiology such as incidence and prevalence of diseases, burden of diseases, quality of life and activity of daily living, medical costs, etc. The insights gained from such data can be extraordinarily valuable (23,24). Indeed, RWD are used to generate real-world evidence, which might be regarded as a "meta-analysis" of accumulated RWD. Increasingly, regulatory authorities are recognising the value of RWD and real-world evidence, especially for rare diseases where it may be practically unfeasible to conduct randomised controlled trials. However, the quality of real-world evidence

depends on the quality of the data collected (26). Experts in the life sciences industry, including pharmaceutical and biotech companies, can use this information to support regulatory approvals and post-approval validations of healthcare products and interventions (27).

Among a wide range of applications, researchers and clinicians have leveraged RWD to evaluate the real-world effectiveness of cancer therapies (28), digital health interventions for mental health (29) and real-word COVID-19 vaccines effectiveness (30).

RWD is expected to play an increasingly important role in healthcare research and decision-making in the years to come (31). However, a major challenge in RWD usage relates to problems with data quality, as RWD that is routinely collected outside of controlled study settings may be inconsistent, inaccurate, or incomplete. The lack of standardisation in data collection and coding across different healthcare systems and regions also poses a challenge for RWD integration and analysis. Ethical and privacy concerns related to patient data collection and sharing are additional aspects for consideration (23,31).

The health insurance data linked to the survey data in this thesis are RWD in their own right. Our linkage has already highlighted a number of challenges with respect to privacy and confidentiality issues. These considerations are likely to be more important when dealing with other types of RWD, such as from wearable devices, the environment, social media, laboratories, etc. While such data can be linked with survey data producing very useful sources of information for researchers and policymakers, this type of linkage also comes with a number of challenges to consider as well as the risk of misuse, such as for commercial purposes. A clear legal framework will therefore need to be established. In addition, as mentioned above, the use of RWD also brings with it technical challenges: the selection of appropriate statistical and epidemiological methods is extremely critical because RWD contain a greater variety of biases, unstructured textual data and the linkage with multiple databases.

8.6. IMPLICATIONS OF THE FINDINGS AND RECOMMENDATIONS

8.6.1. Cross-cutting recommendations

Although the results of this thesis are based on Belgian data, concrete recommendations that go beyond the Belgian context can be made. The following recommendations should be taken into account, bearing in mind that the linkage only concerns data from the health interview survey and health insurance data. The main recommendations are summarised in Table 8.4 and further elaborated below.

Table 8.4: Summary of general recommendations

HIS limitations
• Continue to promote the linkage between databases
Confidentiality and privacy issues
• Work towards a better balance between the right to privacy of respondents and society's right to evidence-based information to improve health • Facilitate access and reuse of data including data linkage
GDPR and consent right
• Gain and maintain the citizens' trust in secondary use of data and data linkage • Adapt the need for consent to the context of the linkages • Improve communication with participants
Data quality
• Plan ahead the linkage of HIS and administrative data
Data substitution
• Substitute HIS information with administrative data as much as appropriate
Contents of linked data
• Extend the linkage to other administrative data sources • Have a good knowledge of data sources and understand the limits of the linked database

8.6.1.1. Continue to promote the linkage between databases

Almost all EU the Member States (MS) organise an HIS among their respective populations. In addition, in the context of universal coverage of the health system in

place in many EU MS (32), MSs have health insurance databases, in some cases for the whole population. The content of these databases may vary from one MS to another. Given the current challenges in primary data collection, such as falling response rates and rising survey costs, it is becoming increasingly urgent to review strategies for collecting population health data effectively. The linkage of existing data is an approach of key importance in this respect as it can generate significant value for public health research, as demonstrated in this thesis.

Data linkage is already well established in several MS, but most often it consists of linking administrative data together (routine administrative linkage). Haneef et al. (2020) conducted a study to describe the current use of individual-level data linkage and artificial intelligence in routine public health activities in European countries and found that the majority of the countries (24 out of 29 respondents) have integrated data linkage into their routine public health activities. Of the 24 countries that practise routine data linkage, 22 link administrative data such as electronic health records, mortality data and specific disease registers, but only 15 link their national health surveys (17). Given the assets of combining administrative health data with primary research data such as those from the national HISs, it seems crucial to promote this type of linkage on a wider scale. This approach can be encouraged, for example, by ensuring that researchers are well informed about the added value of such a linkage, by informing survey participants upstream about the advantages of linkage and the security measures put in place to protect their data, and by encouraging political decision-makers to invest in such a project.

Although this thesis focuses on linkage between HIS and health insurance data, linkage between databases that do not contain HIS data should also be encouraged. For example, a linkage between birth certificates, hospital discharge data and health insurance data has been used to study drug exposure during pregnancy in Belgium (20).

However, in the context of budgetary constraints, it is important for researchers to demonstrate to funders and policymakers the usefulness of such linkage in order to maintain project funding and sustainability and to raise awareness of such initiatives. From a public health perspective, policymakers should continue to invest in data linkages; and the inclusion of other data sources (such as primary-care data and

hospital discharge data) will augment the use of the linked data to expand the evidence base for policymakers and practitioners.

8.6.1.2. Work towards a better balance between the right to privacy of respondents and society's right to evidence-based information to improve health

Privacy considerations must strike a balance between the privacy rights of respondents and society's right to evidence-based information to improve health. Procedures to ensure these considerations need to be optimised in order to be able to carry out data linkages within a reasonable time. For the linkages discussed in this thesis, it took more than a year to complete all the stages, from the preparatory work (meetings with data holders, preparation of the necessary documents and signatures, IRB, etc.) to linkage itself and the availability of the linked data. It is therefore necessary to optimise the overall process in the future.

For researchers wishing to answer research questions requiring data from multiple sources, access to data and carrying out data linkage are often accompanied by extensive applications for data use, data protection concepts and, if necessary, ethics board approval applications – and thus a particularly high workload. This may discourage or significantly delay important research initiatives (33). At European level, in order to unleash the full potential of health data, European Commission presented a regulation to set up the European Health Data Space (EHDS) in November 2020 (33). The creation of a European Data Space, which would include the health sector, is one of the priorities of the Commission for the 2019-2025 period (23). The full proposal for a regulation to set up the EHDS was released in May 2022 (34) and focuses on one hand on the support of the use of health data supporting healthcare delivery, which is called primary use, and secondary use, which is defined as the use of individual-level (personal or non-personal) health data, or aggregated datasets, for the purpose of supporting research, innovation, policy-making, regulatory activities and other uses (35). Data linkage as used in this study falls within this framework of secondary use of health data.

Although the separation principle of linkage and analysis processes is recognised as good practice for protecting confidentiality, allowing linkage and analysis to take place together provides opportunities for both in-depth evaluation of linkage quality and

methodological advances in linkage techniques (36,37). Such an approach is in operation at the Institute for Clinical Evaluative Sciences (ICES) in Ontario. The ICES is legally allowed to receive fully identifiable data in order to perform linkage, to assess data quality and to provide coded data to research staff within the organisation. They operate a hierarchical access policy, which means that only a specific number of people have the highest level of access to all data elements, and most researchers can only access de-identified, coded data relevant to their study (38).

8.6.1.3. Facilitate access and reuse of data including data linkage

A centre for coordinating and linking data could alleviate the problems encountered by researchers and considerably reduce the effort involved in accessing and linking data. In several countries, there are already initiatives relating to the secondary use of data that greatly facilitate the linking of data, such as the Medical Informatics Initiative (Germany), the Health Data Hub (France), the Health Research Infrastructure (Netherlands) and the Personalized Health Network (Switzerland) (23). Although the scope of the data available and the processes for linking and accessing data differ, the guiding principles of these various initiatives are aligned (33). In Belgium, the ongoing initiative of the Belgian Health Data Agency (HDA) will be responsible for facilitating the secondary use of health data and ensuring that this data is reused in a secure and controlled way for health research and innovation. It can support the existing initiatives of the five main federal organisations working with health and healthcare data (Federal Agency for Medicines and Health Products (FAMHP), Federal Public Services Public Health, Food Chain Safety and Environment (FPSHFCSE), Belgian Health Care Knowledge Centre (KCE), National Institute for Health and Disability Insurance (NIHDI) and Sciensano) (23). For now, data linkage is not foreseen as a service in the HDA. However, an active role for the HDA in data linking could greatly facilitate the whole process by reducing the problems associated with it.

8.6.1.4. Gain and maintain the citizens' trust in secondary use of data and data linkage

From a legal perspective, the GDPR considers health data as a special category of personal data whose processing is prohibited other than in exceptional circumstances, such as explicit consent by the data subject or for reasons of public

interest (39). So, with the implementation of the GDPR, the informed consent became mandatory for future planned linkages. Researchers need to put in place strategies to gain the trust of and to involve citizens whose data will be linked (40). The perceived risk to privacy and data confidentiality constitutes one of the primary reasons why respondents decline the linkage requests (41). It is therefore important to emphasise the merits of the research, to stress the importance of altruism (contribution to society) and to address respondents' privacy and confidentiality concerns by informing them of the safeguards put in place to protect their data.

8.6.1.5. Adapt the need for consent to the context of the linkages

For linkages between datasets that already exist, a clear framework of acceptable practices needs to be developed, which the EHDS initiative is attempting to do (42). To maintain population trust in secondary use of data and data linkage, it is imperative that this framework is in line with citizens' values (40). A clear distinction should be made between: 1) Routine linkages, which are usually for primary use and where implicit consent can be assumed because it concerns direct clinical care. However, a harmonised framework needs to be developed in order to streamline secure data flows; 2) Necessary linkages, in a public health crisis, as exemplified by the COVID-19 pandemic and where consent should not be required (43); and 3) Linkages for public health research and surveillance or other scientific research in the public interest, where the preferred legal basis should not be consent, but an explicit legal and ethical framework that is developed by the national health data authorities, resulting in a federated network of Findable, Accessible, Interoperable and Reusable (FAIR), linkable data sources governed by rules that are trusted both by researchers and citizens.

8.6.1.6. Improve communication with participants

The literature suggests a strong correlation between respondents' understanding and how likely they are to give consent (41,44). To achieve higher consent rates, it is necessary to shed light on respondents' understanding of the linkage consent. Several approaches have been proposed to improve linkage consent rates. One of these consists of providing key subgroups that are less likely to understand the linkage request with additional targeted explanatory or informative material. Another approach would be to use tailored messages by asking the consent questions first, then doing

a targeted intervention to address any misunderstandings, before administering the linkage request. It is preferable to ask for linkage consent upfront, which yields higher consent rates (2,3,45).

8.6.1.7. Plan ahead the linkage of HIS and administrative data

A prerequisite for linking HIS and administrative data is the availability of at least one common variable that can be used as linkage key. Linking HIS data with health insurance data is straightforward when a unique personal identifier is available that can be used as linkage key. However, such a linkage becomes complex without a reliable unique identifier. In this situation and to avoid technical problems, it is essential to plan well ahead for the linkage. When researchers plan a new survey, if they have a database with national identifiers, this will be used as the sampling frame. If a linkage is planned, the key must be kept by a TTP. If there is no unique identifier, then the data requirements for probabilistic matching need to be considered in advance. In this case, it is common practice to incorporate and retain personal identifiers (age, gender, date of birth, postcode, etc.) which are also found in the administrative files to which the survey data could be linked. It is worth spending some time on this issue before starting data collection, as it will save a great deal of energy and result in better quality data when the survey data are linked to the administrative records. Consequently, researchers must identify the minimum amount of personal data necessary for the purposes of the linkage.

Furthermore, this thesis has shown that even if the unique personal identifier is used as the linkage key, the risk of missed matches remains. Therefore, it is recommended that deterministic linking be considered alongside probabilistic linking: initial deterministic methods and subsequent probabilistic linking for incomplete links (46,47).

8.6.1.8. Substitute HIS information with administrative data as far as appropriate

In view of the current challenges facing surveys, there is a need to keep survey questionnaires as short as possible. Hence the more information can be obtained through other sources, the shorter can be the questionnaire. When possible, self-reported items should be replaced by administrative data. This will be the case, for example, for cancer screening, reimbursed healthcare use or reimbursed drug use.

However, it is important to keep in mind that the replacement of self-reported information with administrative data can have certain limitations since administrative data have their own shortcomings (e.g. incomplete or missing data, recording errors).

8.6.1.9. Extend the linkage to other administrative data sources

Although this thesis has highlighted the enormous potential of linking HIS data with health insurance data, it would be useful to extend the linkage to other administrative databases with a view to studying specific research questions that otherwise cannot be addressed.

Potential health and health-related administrative data that could be included are hospital discharge data, primary care data, social security data, census data, mortality data, environmental data and disease registries. In this way, the extended linked data would provide a more complete picture of the health and health-related information of the population. The resulting linked data can be used for prospective studies to examine much more precisely the impact of disease, lifestyle and other health determinants on mortality and healthcare consumption, or to estimate the economic impact of disease and ill health. For example, the linkage of HIS with health insurance data and hospital discharge data can be used to study the extent to which individual characteristics and healthcare consumption are associated with hospital outcomes (e.g. complications, mortality). Other examples are provided in the future perspectives section. Therefore, from a public health perspective, policymakers should continue to support systematic data linkages, thereby increasing research opportunities for more evidence-based policies. However, within the context of budgetary constraints, it is important for researchers to demonstrate to funders and policymakers the usefulness of such linkage in order to maintain project funding and sustainability and to raise awareness of such initiatives.

8.6.1.10. Have a good knowledge of data sources and understand the limits of the linked database

Although linked data offers a number of advantages, it is important to bear in mind that the limitations of both data sources remain even after linkage. Researchers need to be aware of this in order to understand and interpret the results with caution.

In addition, given the complexity of administrative data, it is often necessary to involve an expert in these data for the analysis as well as the interpretation of the results. Health administrative data are collected for purposes other than epidemiological research. They are therefore not easy to understand or to use. Expert advice is often needed to make the right choices when planning the analysis.

Moreover, linkage errors are another threat that can significantly distort the results obtained from linked data. Therefore, whatever data linkage methods are used (even if unique personal identifiers are used), researchers should understand linkage errors, evaluate the quality of linked data and validate them. In the event of linkage errors, specific statistical methods should be applied to address them (9,48).

8.6.2. Belgian health interview - specific recommendations

Table 8.5 gives an overview of the recommendations specific to Belgium. Although there may be some overlap between these recommendations and the cross-cutting recommendations above, it is appropriate to mention them here given the specificity of the Belgian context.

Table 8.5: Summary of recommendations specific to Belgium

Confidentiality and privacy issues
• Facilitate overall administrative process
• Optimise the way of dealing with ethical and privacy requirements
Return on investment and content of linked data
• Avoid the 'link and destroy' model • Extend the content of the linked data • Make the linked data available for external users
Complexity of administrative data
• Involve administrative data experts for their advice in any steps of analysis and interpretation of the findings.

8.6.2.1. Facilitate overall administrative process

For the organisation of the BHIS, ethical approval was obtained from the EC of Ghent University/University Hospital. Approval of the actual linkage procedure was obtained

from the ISC which acts as IRB. The ISC approval process is usually complex and time-consuming. For both HISlink 2013 and HISlink 2018, it took several months to get the ISC approval. A need to facilitate the establishment of linkages has also been highlighted in Sciensano 's recent study which investigated opportunities for a population-based cohort in Belgium (49). Therefore, to facilitate data linkage and overcome lengthy negotiations and ad hoc approval processes for each BHIS-BCHI linkage, it would be useful to set up a kind of "umbrella" agreement protocol for more structural linkage for public institutions such as Sciensano. Such a "umbrella" agreement protocol could cover multiple years and multiple waves of BHIS-BCHI linkages. However, the rights of participants have to be ensured. In addition, a tool such as the new Belgian HDA should also help to facilitate the entire administrative process.

8.6.2.2. Optimise the way of dealing with ethical and privacy requirements

An important challenge when linking BHIS data with other databases is the privacy of the respondents. Many topics addressed in BHIS are sensitive and linkages with administrative databases will increase the risk of identification. To preserve privacy and prevent the disclosure of sensitive information, a Data Protection Impact Assessment (DPIA) as well as the SCRA was conducted, and the linkage was carried out by a Trusted Third Party (TTP) to comply with the separation principle of linkage and analysis processes. The separation principle means that those conducting the linkage (often TTPs) only have access to a set of identifiers, whilst those analysing the linked data only have access to de-identified attribute data. Although this approach is considered as good practice, it causes a significant delay in the linkage process due to the administrative steps that take time (e.g. the signature of an official agreement between the parties involved). Researchers should consider the ISC approval process in the timeline of the whole project. Researchers can also get together with all the partners to discuss the barriers that are making the process so long and look at how it can be optimised in terms of timing. Furthermore, although the separation principle reduces the risk of disclosure of sensitive information about individuals, it means that important aspects of the linkage process are obscured, which makes it difficult for researchers to judge the reliability of the resulting linked data for their required purposes. Consequently, allowing linkage and analysis to take

place at the same time, or at least involving some of the researchers in the linkage processes, would be a better approach (36,37).

To comply with the GDPR, an effective opt-in linkage consent form has to be signed by the participants. HISlink 2013 and 2018 were not consent-based (exemptions, linkage planned before the implementation of the GDPR). As from BHIS 2023 onwards, a signed informed consent for data linkage from BHIS participants is required to link their BHIS and BCHI data. However, requiring explicit consent leads to an additional burden on respondents and interviewers, as well as to a potential consent bias. A legal framework to facilitate linkages between existing data sources should be put in place, which requires a global reflection and approach with all parties concerned. For example, there is a legal framework that allows Statbel to link databases without the formal consent of individuals. But here too, privacy considerations are very important. The objectives of the HDA are entirely relevant in this respect. In collaboration with data providers and data users, it will contribute to making health data, in all its facets, more easily, uniformly and transparently available in a secure environment. The HDA aims to encourage population management, scientific research, innovation and policy development to improve the health of citizens. Ultimately, the Belgian HDA will facilitate the secondary use of health data including data linkage (50).

8.6.2.3. Avoid the 'link and destroy' model

The HISlink is based on the 'link and destroy' philosophy involving a limited data retention time. Indeed, the linked data are only available to researchers until the end of a predetermined retention date explicitly indicated in the ISC approval. After this period, the linked data are destroyed. As a result, the return on investment in linked data may be limited. The data retention can be extended upon a request for amendment of the ISC approval. In addition, 'link and destroy' has an impact on the availability and accessibility of data for research and policy development (51). From a broader perspective, it would be useful to have streamlined approval processes for efficient data access. Indeed, some jurisdictions adopt approaches for timely and cost-effective access to linked data (e.g. those in Ontario, Wales and Australia where linkage keys can be held in perpetuity), while others such as in Belgium are restricted by the 'link and destroy' model. In the later model, linked data cannot be reused or are

destroyed after a predefined data-retention time. A clear data use agreement for governmental institutions, administrations and universities allowing share and use of the linked databases for at least several years (or even in perpetuity) in a secure manner. Such strategies would allow exploiting the full potential of the linked data. But other researchers should be aware of the existence of these linked databases in order to use them. Also, data should then be thoroughly documented with clear metadata in order to be re-usable by a large community of researchers. In addition, keeping the linked data beyond the retention date could help to perform specific analysis requiring high sample size using an accelerated longitudinal design analysis (putting together linked data of several BHIS waves).

8.6.2.4. Extend the content of the linked data

There is an enormous amount of unlinked data sets (surveys, cohorts, administrative databases), all of which are valuable in their own right. However, researchers are increasingly faced with research questions requiring more comprehensive data. For practical reasons and budgetary constraints, survey data cannot contain all the relevant and detailed information that a researcher might need, for example, detailed information on the use of healthcare services or their cost. In addition, specific hard-to-reach population subgroups, such as people of lower socio-economic status, are under-represented in surveys, as are rare events such as low-prevalence diseases for example due to sample size limitations. On the other hand, administrative data are very rich and detailed but lack crucial information such as socio-economic information and health determinants for further analysis. To answer complex research questions requiring more comprehensive information, it is worth investing in the linkage of several data sources rather than undertaking several primary data collections despite all the effort involved.

In Belgium, in addition to BHIS and BCHI data, there are many high-quality administrative data sources. These administrative data sources provide valuable routine information on the health and health-related topics of the Belgian population. They include data on hospital discharges, mortality, social security and the labour market, the environment, etc. From a research point of view, extending the data currently linked to these administrative data will make it possible to answer certain crucial health questions that are difficult or even impossible to answer. For example,

the inclusion of hospital discharge data could help target internal quality improvement efforts for specific patient groups (e.g. preventive care for diabetics) or help assess the determinants of hospitalisation and understand the underlying factors that influence the length of hospitalisation and the costs of illness. Mortality data is also important to consider. Extending HISlink data to mortality data will make it possible to study cause-specific mortality as a function of socio-economic status.

Furthermore, understanding the factors that determine the healthcare-seeking behaviour of the population is extremely important for the development of rational policy aimed at providing accessible, efficient and (cost-effective) services. Factors influencing health-seeking behaviour include socio-demographic factors, physical accessibility, but also the family environment and social support. In this context, data obtained through administrative processes such as the social security system are essential for complementing (longitudinal) health survey data and for identifying direct and indirect effects.

Another example concerns participation in the labour market. Labour market participation is an important policy objective, and substantial resources are made available by the federal government to reduce poverty by increasing the participation rate. Important barriers and resources have an impact on labour market participation, including health and disability, geographical differences, educational attainment and household characteristics, as well as well-being and coping skills, social support and self-efficacy. In order to assess their potential impact on sustainable labour market participation and the reduction of inequalities, data on these topics should be brought together to explore their complex interactions and unravel the complex determinants of labour market participation

Since BHIS and BCHI represent important sources of population-based data, many linkage projects involve either BHIS or BCHI data or both, sometimes in addition to supplementary data sources such as cancer registry data, mortality data, etc. Hence, the BHIS-BCHI data sources could act as a spine where other data sources (nodes) could be linked on a project basis. A similar approach is implemented in New Zealand, the Integrated Data Infrastructure (IDI). The IDI consists of a central spine and many nodes (collections of datasets linked to the spine). For example, the health node includes datasets such as survey data, pharmaceutical dispensing, lab tests, and hospital discharges, all linked by the National Health Index (NHI) (9). Although this

approach cannot be reproduced as it stands in the Belgian context, it can be adapted to ensure the sustainability of HISLink and make better use of linked datasets, exploiting its full potential and reducing the duration of the linking process. However, an important disadvantage of maintaining linked BHIS-BCHI data as a spine is that involving HIS data reduces the sample size considerably for rare events, making, for example, a BHIS - cancer registry linkage less interesting, whereas a BCHI cancer registry linkage may be of interest. Therefore, depending on the research question and the information needed, the spine-node approach should be adapted.

8.6.2.5. Make the linked data available for external users

Due to legal constraints, access to HISlink data is currently restricted to Sciensano researchers who are registered with the IMA as users of HISlink data. Although more and more researchers are currently using HISlink data within Sciensano, it would be beneficial to consider making de-identified data available to external public health researchers on approval. Indeed, to further leverage the linked data, the owners of the data, i.e. Sciensano, the IMA and the sponsor (NIHDI), could retain ownership but make the data available to other research studies in line with the primary purpose of HISlink, subject to the owners' approval. An example of such an approach in cancer research is the National Cancer Institute's (NCI's) linked Surveillance, Epidemiology and End Results (SEER) Medicare files, where the NCI retains ownership of the data and makes it available to approved research studies that guarantee patient and provider confidentiality in SEER areas (52).

Currently, researchers who wish to link BHIS and BCHI data have to go through the entire procedures (for example IRB approval). The possibility of this linkage should be more widely spread and researchers who want to do this type of linkage given assistance.

8.6.2.6. Involve administrative data experts for their advice in any steps of analysis and interpretation of the findings

Administrative data are complex and not initially intended for epidemiological research. They are also conceptually different from survey data. Therefore, to avoid errors and misinterpretation of results, it is important to call on specific experts to

provide advice at each stage of the research conceptualisation, analysis and interpretation of results. Good collaboration between all the partners involved is also important.

8.7. FINAL CONCLUSIONS

Data linkage brings significant added value to public health researchers. It has made it possible to answer policy-relevant research questions that cannot be answered using separate tools alone. It has also made it possible to assess the validity of HIS data in relation to health insurance administrative data. From a public health perspective, policymakers should continue investing in data linkages and the inclusion of other data sources such as hospital discharge data, mortality data, and primary healthcare data. Linking different administrative data sources will augment the use of the linked data to expand the evidence base for policymakers and practitioners and can thus enrich population-based surveillance and the field of research into public health. Considering the strengths and limitations of different data sources, the opportunity to link multiple data sources could potentially enable a wider range of research questions to be addressed. However, linking survey data to administrative data is not without its challenges and these have to be tackled. Although some aspects of the HISlink may be specific to the Belgian context, we believe that the results of this thesis have much broader implications and could be useful to researchers who plan to link health survey data with health administrative data for their respective projects.

8.8. BIBLIOGRAPHY

1. Sediq R, Van Der Schans J, Dotinga A, Alingh RA, Wilffert B, Bos JH, et al. Concordance assessment of self-reported medication use in the Netherlands three-generation Lifelines Cohort study with the pharmacy database iaDB. nl: The PharmLines initiative. Clinical epidemiology. 2018;981–9.
2. van Brug HE, Rosendaal FR, van Steenbergen LN, Nelissen RG, Gademan MG. Data linkage of two national databases: Lessons learned from linking the Dutch Arthroplasty Register with the Dutch Foundation for Pharmaceutical Statistics. Plos one. 2023;18(3):e0282519.
3. Applying for linked data from the Lifelines Cohort Study and IADB.nl database [Internet]. 2021. Available from: http://wiki-lifelines.web.rug.nl/lib/exe/fetch.php?media=pharmlines_procedures_20210415.pdf
4. Van der Heyden J, Charafeddine R, De Bacquer D, Tafforeau J, Van Herck K. Regional differences in the validity of self-reported use of health care in Belgium: selection versus reporting bias. BMC Med Res Methodol. 2016 Dec;16(1):98.
5. Tolonen H, Kopra K, Vartiainen E. The effect of non-participation on the estimation of smoking trends: Hanna Tolonen. The European Journal of Public Health. 2015;25(suppl_3):ckv175-050.
6. Calderwood L, Lessof C. Enhancing Longitudinal Surveys by Linking to Administrative Data. In: Lynn P, editor. Methodology of Longitudinal Surveys [Internet]. Chichester, UK: John Wiley & Sons, Ltd; 2009 [cited 2020 Sep 29]. p. 55–72. Available from: http://doi.wiley.com/10.1002/9780470743874.ch4
7. Drieskens S, Demarest S, Bel S, De Ridder K, Tafforeau J. Correction of self-reported BMI based on objective measurements: a Belgian experience. Arch Public Health. 2018 Dec;76(1):10.
8. Dutton DJ, McLaren L. The usefulness of “corrected” body mass index vs. self-reported body mass index: comparing the population distributions, sensitivity, specificity, and predictive utility of three correction equations using Canadian population-based data. BMC public health. 2014;14(1):1–11.
9. Kvalsvig A, Gibb S, Teng A. Linkage error and linkage bias: A guide for IDI users. University of Otago. 2019.
10. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. International Journal of Epidemiology. 2017 Oct 1;46(5):1699–710.
11. Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. Annals of Human Biology. 2020 Feb 17;47(2):218–26.
12. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. International Journal of Epidemiology. 2014 Dec 1;43(6):1969–85.

13. Lamers LM, van Vliet RCJA. The Pharmacy-based Cost Group model: validating and adjusting the classification of medications for chronic conditions to the Dutch situation. Health Policy. 2004 Apr;68(1):113–21.
14. Canney M, McNicholas T, Scarlett S, Briggs R. Prevalence and Impact of Chronic Debilitating Disorders. In: MCGARRIGLE C, DONOGHUE O, SCARLETT S & KENNY R (eds) Health and Wellbeing: Active Ageing for Older Adults in Ireland - Evidence from the Longitudinal Study on Ageing [Internet]. Dublin: TILDA.; 2017. Available from: https://www.doi.org/10.38018/TildaRe.2017-01.c7
15. Knies G, Burton J, Sala E. Consenting to health record linkage: evidence from a multi-purpose longitudinal survey of a general population. BMC Health Services Research. 2012;12(1):1–6.
16. Harron K. Data linkage in medical research. bmjmed. 2022 Mar;1(1):e000087.
17. Haneef R, Delnord M, Vernay M, Bauchet E, Gaidelyte R, Van Oyen H, et al. Innovative use of data sources: a cross-sectional study of data linkage and artificial intelligence practices across European countries. Arch Public Health. 2020 Dec;78(1):55.
18. Ward A, Sarraju A, Lee D, Bhasin K, Gad S, Beetel R, et al. COVID-19 is associated with higher risk of venous thrombosis, but not arterial thrombosis, compared with influenza: Insights from a large US cohort. PLoS One. 2022;17(1):e0261786.
19. Canova C, Cantarutti A. Population-based birth cohort studies in epidemiology. International journal of environmental research and public health. 2020;17(15):5276.
20. Larcin L, Neven A, Damase-Michel C, Kirakoya-Samadoulougou F. Belgian medication exposure during pregnancy (BeMeP), a new nationwide linked database: Linkage methods and prevalence of medication use. Pharmacoepidemiology and drug safety. 2023.
21. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Medical Research Methodology. 2022;22(1):287.
22. FDA US Food and Drug Administration. Real-World Evidence [Internet]. 2022 [cited 2023 Dec 7]. Available from: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
23. Maes I, Kok E, Dewulf G. Recommendations on a Real-World Data Strategy for Belgium [Internet]. 2022 [cited 2023 Sep 21]. Available from: https://www.inovigate.com/media/filer_public/36/dd/36ddcd12-564e-4678-9388-053d7adc6b12/report_recommendations_on_rwd_strategy_for_belgium_final_template_used.pdf

24. Togo K, Yonemoto N. Real world data and data science in medical research: present and future. Japanese Journal of Statistics and Data Science. 2022;5(2):769–81.
25. Bourke A, Dixon WG, Roddam A, Lin KJ, Hall GC, Curtis JR, et al. Incorporating patient generated health data into pharmacoepidemiological research. Pharmacoepidemiology and Drug Safety. 2020;29(12):1540–9.
26. Demetri GD, Stacchiotti S. Contributions of real-world evidence and real-world data to decision-making in the management of soft tissue sarcomas. Oncology. 2021;99(1):3–7.
27. Jones G. Real-world data (RWD) vs. real-world evidence (RWE) [Internet]. 2022 [cited 2023 Dec 7]. Available from: https://blogs.oracle.com/life-sciences/post/real-world-data-vs-real-world-evidence
28. Phillips CM, Parmar A, Guo H, Schwartz D, Isaranuwatchai W, Beca J, et al. Assessing the efficacy-effectiveness gap for cancer therapies: a comparison of overall survival and toxicity between clinical trial and population-based, real-world data for contemporary parenteral cancer therapeutics. Cancer. 2020;126(8):1717–26.
29. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR mHealth and uHealth. 2018;6(11):e12106.
30. Henry DA, Jones MA, Stehlik P, Glasziou PP. Effectiveness of COVID-19 vaccines: findings from real world studies. The Medical Journal of Australia. 2021;215(4):149.
31. Mondragon LG. Real World Data: Opening New Avenues for Health Research [Internet]. 2023. Available from: https://www.the-scientist.com/sponsored-article/real-world-data-opening-new-avenues-for-health-research-71090#:~:text=Real%20world%20data%20can%20be,make%20discoveries%20in%20health%20research.
32. OECD, European Union. Health at a Glance: Europe 2022: State of Health in the EU Cycle [Internet]. OECD; 2022 [cited 2023 Aug 10]. (Health at a Glance: Europe). Available from: https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-europe-2022_507433b0-en
33. Panteli D, Polin K, Webb E, Allin S, Barnes A, Degelsegger-Márquez A, et al. Health and care data: approaches to data linkage for evidence-informed policy. 2023.
34. European Commission communication-from-the-commission-to-the-european-parliament-and-the-council. A European Health Data Space: harnessing the power of health data for people, patients and innovation [Internet]. Koninklijke Brill NV; [cited 2023 Aug 30]. Available from: https://primarysources.brillonline.com/browse/human-rights-documents-

online/communication-from-the-commission-to-the-european-parliament-and-the-council;hrdhrd46790058

35. Marcus JS, Martens B, Carugati C, Bucher A, Godlovitch I. The European Health Data Space. SSRN Journal [Internet]. 2022 [cited 2023 Aug 30]; Available from: https://www.ssrn.com/abstract=4300393
36. Aldridge RW, Shaji K, Hayward AC, Abubakar I. Accuracy of Probabilistic Linkage Using the Enhanced Matching System for Public Health and Epidemiological Studies. Pacheco AG, editor. PLoS ONE. 2015 Aug 24;10(8):e0136179.
37. Harron K, Goldstein H, Wade A, Muller-Pebody B, Parslow R, Gilbert R. Linkage, Evaluation and Analysis of National Electronic Healthcare Data: Application to Providing Enhanced Blood-Stream Infection Surveillance in Paediatric Intensive Care. Trotter CL, editor. PLoS ONE. 2013 Dec 20;8(12):e85278.
38. Harron K, Goldstein H, Dibben C. Methodological developments in data linkage. John Wiley & Sons; 2015.
39. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [Internet]. 2016 [cited 2023 Sep 28]. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&qid=1695900759326
40. Maddocks J, Mathieu L, Richards R, Saelaert M, Van Hoof W. tehdas-healthy-data-an-online-citizen-consultation-about-health-data-reuse-intermediate-report.pdf [Internet]. 2022 Jun [cited 2022 Sep 27]. Available from: https://tehdas.eu/app/uploads/2022/07/tehdas-healthy-data-an-online-citizen-consultation-about-health-data-reuse-intermediate-report.pdf
41. Sakshaug JW, Schmucker A, Kreuter F, Couper MP, Holtmann L. Respondent understanding of data linkage consent. Survey Methods: Insights from the Field (SMIF). 2021.
42. European Commission. REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space [Internet]. 2022 [cited 2023 Feb 9]. Available from: https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF
43. McLennan S, Celi LA, Buyx A. COVID-19: Putting the General Data Protection Regulation to the Test. JMIR Public Health Surveill. 2020 May 29;6(2):e19279.
44. Marie Thornby LC. Collecting Multiple Data Linkage Consents in a Mixed-mode Survey: Evidence from a large-scale longitudinal study in the UK. 2018 [cited 2022 Sep 27]; Available from: https://surveyinsights.org/?p=9734
45. March S. Individual Data Linkage of Survey Data with Claims Data in Germany—An Overview Based on a Cohort Study. IJERPH. 2017 Dec 9;14(12):1543.

46. Jamieson E, Roberts J, Browne G. The feasibility and accuracy of anonymized record linkage to estimate shared clientele among three health and social service agencies. Methods of information in medicine. 1995;34(04):371–7.
47. Berkeley M. TEXTBOOK OF MEDICAL RECORD LINKAGE. The Journal of the Royal College of General Practitioners. 1987;37(304):518.
48. Fawcett J, Blakely T, Atkinson J. Weighting the 81, 86, 91 & 96 census-mortality cohorts to adjust for linkage bias. Department of Public Health, Wellington School of Medicine and Health …; 2002.
49. Schutte N, Saelaert M, Bogaert P, De Ridder K, Van Oyen H, Van der Heyden J, et al. Opportunities for a population-based cohort in Belgium. Arch Public Health. 2022 Aug 11;80(1):188.
50. Constandt (Hans. Towards the establishment of the Belgian Health Data Authority. LCD Study day “health information and policy; 2022 Oct 6; Sciensano, Brussels, Belgium.
51. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. Big Data & Society. 2017 Dec;4(2):205395171774567.
52. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future: Health Services Research and Data Linkages. Health Services Research. 2010 Oct;45(5p2):1468–88.

Acknowledgement

One day, Stefaan asked me the question that led to this work: are you interested in doing a PhD in addition to your work? I answered YES without hesitation. I couldn't let this great opportunity pass me by. But, at the time, I didn't really know what I was getting myself into, in all honesty. Another day, when I was discussing my first article, Herman said to me: «…that's good what you've done, but you still need to do this and that.... ». I replied «...that's difficult, I've got lots of other things to do». He said: «…it's possible, you've got two hands and several fingers, so you can do several things at once». I realised then that I'd have to work hard to get there. And here I am, a little proud, but very grateful to be able to complete this chapter.

First of all, I would like to thank my promotors, Prof. dr. Olivier Bruyère and Prof. dr. Herman Van Oyen. I have been fortunate to benefit from your advice, support and confidence throughout the process and I have greatly appreciated your invaluable contribution.

Thanks to the members of my jury, Prof. dr. Olivier Bruyère, Prof. dr. Herman Van Oyen, dr. Johan Van der Heyden, Prof. dr. Katrien Vanthomme, Prof. dr. Olivier Ethgen, Prof. dr. Pierre Gillet and Prof. dr. Romana Haneef for their critical reading of my thesis and their constructive comments. My thanks also go to the members of my informal doctoral committee: Prof. dr. Olivier Bruyère, Prof. dr. Herman Van Oyen, dr. Johan Van der Heyden, Prof. dr. Jean-Yves Reginster, Prof. dr. Olivier Ethgen, Prof. dr. Pierre Gillet. It was an honour to receive annual advice from you.

I would also like to thank all members of HIS team: Stefaan, Lydia, Sabine, Rana, Elise, Johan, Jean (our former head of service), Pierre, Lize, Gwendoline, Helena, Christina and Manon for the good team spirit and a good working atmosphere. Thank you Stefaan, for the opportunities you have given me, for your confidence, your support and for always believing that I would succeed in this project, even when I couldn't see the light at the end of it. My special thanks go to Johan. I wanted to express my personal gratitude for the effort and extra time you have contributed to this work. The HISlink project is not a ‘cakewalk’. I really appreciated your support, guidance and encouragement. Throughout all the process I have been able to count on your critical advice and valuable input. Rana, you also deserves a heartfelt thank

you for all the support, input, guidance and motivation you have given me. Thank you Stefaan, Lydia, Sabine, Rana, Elise, Johan and Lize for reading through some parts of this thesis. Thanks also to Lydia for proofreading several articles in English.

Thank you to all the co-authors of the articles that contributed to this thesis. Your advice and critical reflections and discussions have all contributed to the successful completion of this thesis.

Thanks to Nienke and Marlies for their help with the discussion section.

I would like to thank Statbel, the Belgian statistical office, which was responsible for HIS sample selection and fieldwork management. Thanks to Statbel and the InterMutualistic Agency (IMA) for their involvement in the process of data linkage. And of course, all the people who voluntarily participated in the Belgian health interview survey. Without them this research would not have been possible.

I also really appreciated Ledia's administrative help with the layout of this thesis. Thank you very much!

I would also like to express my sincere thanks to my parents. I am little sad that you're no longer here to celebrate this stage in my life. Thanks to my sisters and brother for their support and motivation in my moments of doubt.

Finally, I would like to thank my children, Ousmane and Almamy, my biggest fans, for their understanding when I was less available. Vous voulez faire comme maman, alors vous savez ce qui vous attend desormais ! Mamady, thank you for your constant support, your understanding and for taking care of the boys when I was less available. I could not have done it without you!

This PhD was written in the context of the HISlink project, project financed by NIHDI.

December 7, 2023

Finaba Berete

About the author

Finaba Berete graduated as a Medical Doctor from the Gamal Abdel Nasser University in Conakry (Guinea) in 2005. After obtaining a Master's degree in Public Health, Epidemiology and Health Economics (with honours) from the University of Liège in 2010, she worked as a project officer with the non-profit organisation Cap Santé from 2011 to 2013 and was in charge of the management, planning and monitoring of health projects in the DRC. Always passionate about research, she took an intensive training in data analysis and biostatistics in R&D, at the GIGA biotechnology center, Liège in 2014-2015, followed by an internship in biostatistics department of University hospital of Liège. She was then recruited by Health Observatory of the Province of Namur where she was involved in the conceptual work of the health surveys, drafting the protocol, designing the questionnaire, carrying out the statistical analyses, writing the report and presenting the results.

In 2016, she joined Sciensano (formerly WIV-ISP). Initially she worked on the Autoweb project which consisted to explore item non-response in self-administrated questionnaire of the BHIS. Since 2017, she is responsible of the HISlink project in which her main activities consist of requesting authorisation demand for the linkage, project monitoring, data analysis, reports writing and dissemination of the results. To ensure the HISlink project runs smoothly, she has completed intensive training in 'Introductory Analysis of Linked Health Data' (2019, 35 professional development hours) and in 'Advanced Analysis of Linked Health Data' (2023, 35 professional development hours) at Swansea University (UK) and the University of Western Australia. She is also involved in BHIS, where she is in charge of the cancer screening and vaccination modules. Since 2022, she has been Section Editor for Archives of Public Health (Methodology Section).

Publications list

Articles in international peer-reviewed journals included in Science Citation Index (A1), including articles accepted for publication, as well as submitted manuscripts, chronologically ordered

- **Publications contribute to this thesis**

Berete F, Gisle L, Demarest S, Charafeddine R, Bruyère O, Van den Broucke S, Van der Heyden J. Does health literacy mediate the relationship between socioeconomic status and health(-related) outcomes in the Belgian adult population? Submitted to BMC Public Health.

Berete F, Demarest S, Charafeddine R, De Ridder K ,Van Oyen H, Van Hoof W, Bruyère O, Van der Heyden J. Linking health survey data with health insurance data: methodology, challenges, opportunities and recommendations for public health research. An experience from the HISlink project in Belgium. Archives of Public Health 81.1 (2023): 198.

Berete F, Demarest S, Charafeddine R, Ridder K, Vanoverloop J, Oyen H, et al. Predictors of Nursing Home Admission in the Older Population in Belgium. BMC Geriatrics 22.1 (2022): 1-13.

Van der Heyden J, **Berete F**, Renard F, Vanoverloop J, Devleesschauwer B, De Ridder K, et al. Assessing polypharmacy in the older population: Comparison of a self-reported and prescription-based method. Pharmacoepidemiol Drug Saf. 2021 Dec;30(12):1716–26

Berete F, Van der Heyden J, Demarest S, Charafeddine R, Tafforeau J, Van Oyen H, et al. Validity of self-reported mammography uptake in the Belgian health interview survey: selection and reporting bias. European Journal of Public Health. 2021 Feb 1;31(1):214–20.

Berete F, Demarest S, Charafeddine R, Bruyère O, Van der Heyden J. Comparing health insurance data and health interview survey data for ascertaining chronic disease prevalence in Belgium. Arch Public Health. 2020 Dec;78(1):120

- **Other publications**

Bich Tran P, Van Olmen J**, Berete F**, Gorasso V, Van der Heyden J, Willem L, Loss of productivity and indirect cost of multimorbidity in Belgium. To be submitted to: TBD.

Bich Tran P, Nikolaidis G.F, Abatih E, Bos P, **Berete F**, Gorasso V, Van der Heyden J, Kazibwe J, Tomeny E.M, Van Hal G, Beutels P, Van Olmen J. Multimorbidity Healthcare Expenditure in Belgium: A Four-Year Analysis (COMORB study). Submitted to Health Affairs

Demarest, S.; **Berete, F**.; Baeyens, Y.; Molenberghs, G.; Drieskens, S.; Charafeddine, R.; Braekman, E.; Van Oyen, H.; Van Hal, G. Does Field Substitution Impact the Educational Profile of the Belgian Health Interview Survey Net Sample? Submitted to Plos One.

Demoury C, Aerts R, **Berete F**, Lefebvre W, Pauwels A, Vanpoucke C, Van der Heyden J, De Clercq. EM. Impact of short-term exposure to air pollution on natural mortality and vulnerable populations: a multi-city case-crossover analysis in Belgium. Submitted to Environmental Pollution.

Vasquez, M. S., Mertens, E., **Berete, F**., Van der Heyden, J., Peñalvo, J. L., & Vandevijvere, S. Comparing self-reported health interview survey and pharmacy billing data in determining the prevalence of diabetes, hypertension, and hypercholesterolemia in Belgium. Archives of Public Health, 81(1) (2023):121.

Demoury, C., De Troeyer, K., **Berete, F**., Aerts, R., Van Schaeybroeck, B., Van der Heyden, J., & De Clercq, E. M. Association between temperature and natural mortality in Belgium: Effect modification by individual characteristics and residential environment. *Science of the Total Environment* 851 (2022): 158336.

Braekman, E.; Charafeddine, R.; **Berete, F**.; Bruggeman, H.; Drieskens, S.; Gisle, L.; Hermans, L.; Van der Heyden, J.; Demarest, S. Data collection in pandemic times: the case of the Belgian COVID-19 Health surveys. Arch Public Health 81, 124 (2023). https://doi.org/10.1186/s13690-023-01135-x.

Demarest, S.; Molenberghs, G.; **Berete, F**.; Charafeddine, R.; Van Oyen, H.; Van Hal, G. Time Trends in the Use of Field-Substitution in the Belgian Health Interview Survey. Archives of Public Health 2022, 80 (1), 229. https://doi.org/10.1186/s13690-022-00982-4.

Bruggeman, H.; Smith, P.; **Berete, F**.; Demarest, S.; Hermans, L.; Braekman, E.; Charafeddine, R.; Drieskens, S.; De Ridder, K.; Gisle, L. Anxiety and Depression in Belgium during the First 15 Months of the COVID-19 Pandemic: A Longitudinal Study. Behavioral Sciences 2022, 12 (5).

Vandevijvere, S.; De Ridder, K.; Drieskens, S.; Charafeddine, R.; **Berete, F**.; Demarest, S. Food Insecurity and Its Association with Changes in Nutritional Habits among Adults during the COVID-19 Confinement Measures in Belgium. Public Health Nutrition 2021, 24 (5), 950–956

Braekman, E.; Demarest, S.; Charafeddine, R.; **Berete, F**.; Drieskens, S.; Van der Heyden, J.; Van Hal, G. Response Patterns in the Belgian Health Interview Survey: Web versus Face-to-Face Mode. European Journal of Public Health 2020, 30.

Nguyen, D.; Hautekiet, P.; **Berete, F**.; Braekman, E.; Charafeddine, R.; Demarest, S.; Drieskens, S.; Gisle, L.; Hermans, L.; Tafforeau, J.; Van der Heyden, J. The Belgian Health Examination Survey: Objectives, Design and Methods. Archives of Public Health 2020, 78 (1), 50. https://doi.org/10.1186/s13690-020-00428-9.

Braekman, E.; Charafeddine, R.; Demarest, S.; Drieskens, S.; **Berete, F**.; Gisle, L.; Van der Heyden, J.; Van Hal, G. Comparing Web-Based versus Face-to-Face and Paper-and-Pencil Questionnaire Data Collected through Two Belgian Health Surveys. International Journal of Public Health 2020, 65 (1), 5–16. https://doi.org/10.1007/s00038-019-01327-9.

Braekman, E.; Drieskens, S.; Charafeddine, R.; Demarest, S.; **Berete, F**.; Gisle, L.; Tafforeau, J.; Van der Heyden, J.; Van Hal, G. Mixing Mixed-Mode Designs in a National Health Interview Survey: A Pilot Study to Assess the Impact on the Self-Administered Questionnaire Non-Response. BMC Medical Research Methodology 2019, 19. https://doi.org/10.1186/s12874-019-0860-3.

Berete, F.; Van der Heyden, J.; Demarest, S.; Charafeddine, R.; Gisle, L.; Braekman, E.; Tafforeau, J.;Molenberghs, G. Determinants of Unit Nonresponse in Multi-Mode Data Collection: A Multilevel Analysis. PLOS ONE 2019, 14, e0215652. https://doi.org/10.1371/journal.pone.0215652.

Braekman, E.; **Berete, F**.; Charafeddine, R.; Demarest, S.; Drieskens, S.; Gisle, L.; Molenberghs, G.; Tafforeau, J.; Van der Heyden, J.; Van Hal, G. Measurement Agreement of the Self-Administered Questionnaire of the Belgian Health Interview Survey: Paper-and-Pencil versus Web-Based Mode. PLOS ONE 2018, 13, e0197434. https://doi.org/10.1371/journal.pone.0197434

National reports

Berete F, Van der Heyden J, Charafeddine R, Demarest S. HISlink - Mediating effect of health literacy on the relationship between socioeconomic status and health(-related) outcomes. Brussels, Belgium: Sciensano. Report number: D/2023.14.440/48 (2023). https://doi.org/10.25608/acck-nb22

Gisle L, **Berete F**, Braekman E, Bruggeman H, Charafeddine R, Demarest S, Drieskens S, Flamant N, Hermans L, Nélis G, Smith P. Sciensano. Dixième enquête de santé COVID-19 : Résultats préliminaires. Bruxelles, Belgique. Avril 2022 ; Numéro de dépôt : D/2022/14.440/18. Disponible en ligne: https://doi.org/10.25608/mve9-bk51

Gisle L, **Berete F**, Braekman E, Charafeddine R, Demarest S, Drieskens S, Hermans L, Nélis G, Van der Heyden J. Sciensano. Neuvième enquête de santé COVID-19 : Résultats préliminaires. Bruxelles, Belgique. Janvier 2022 ; Numéro de dépôt : D/2022/14.440/3. Disponible en ligne: https://doi.org/10.25608/evrs-je22

Berete F, Van der Heyden J, Demarest S, Charafeddine R. Couplage des données de l'enquête de santé avec les données des organismes assureurs – Hislink 2013. Étude comparative sur la prévalence des maladies chroniques : analyses supplémentaires. Bruxelles, Belgique : Sciensano ; Numéro de dépôt : D/2021/14.440/40 (2021).

Berete F, Vander Heyden J, Demarest S, Charafeddine R. Couplage des données de l'enquête de santé avec les données des organismes assureurs - HISlink 2013 Méthodologie et étude comparative sur la prévalence des maladies chroniques. Bruxelles, Belgique : Numéro de rapport : D/2021/14.440.39 .

Demarest S, Charafeddine R, **Berete F**, Braekman E, Bruggeman H, Drieskens S, Gisle L, Hermans L, Leclercq V, Van der Heyden J. Sciensano. Sixième enquête de santé COVID-19. Bruxelles, Belgique ; Numéro de dépôt :D/2021/14.440.30. https://doi.org/10.25608/j877-kf56

Gisle L, **Berete F**, Braekman E, Bruggeman H, Charafeddine R, Demarest S, Drieskens S, Van der Heyden J. Sciensano. Septième enquête de santé COVID-19 - Résultats. Bruxelles, Belgique. Septembre 2021 ; Numéro de dépôt : D/2021/14.440.50. https://doi.org/10.25608/ht7a-8923

Charafeddine R, **Berete** F, Braekman E, Bruggeman H, Demarest S, Drieskens S, Gisle L, Nélis. G. Sciensano. Huitième enquête de santé COVID-19. Bruxelles, Belgique ; Numéro de dépôt : D/2021/14.440/82 Disponible en ligne:: https://doi.org/10.25608/hqy9-m065.

Van der Heyden J, **Berete F**, Drieskens S. Enquête de santé. 2018. Soins ambulatoires dispensés par les médecins et les dentistes. Bruxelles, Belgique : Sciensano ; Numéro de rapport : D/2020/14.440/18. Disponible en ligne : www.enquetesante.be

Drieskens F, **Berete F**, Gisle L. Enquête de santé 2018 : Contacts avec des services paramédicaux. Bruxelles, Belgique : Sciensano ; D/2020/14.440/20. Disponible en ligne : www.enquetesante.be

Driesken S, **Berete F**, Scohy A. Enquête de santé 2018 : Contacts avec des prestataires de thérapies non-conventionnelles. Bruxelles, Belgique : Sciensano ; D/2020/14.440/21. Disponible en ligne : www.enquêtesante.be

Charafeddine R, Demarest S, **Berete F**, Drieskens S. Enquête de santé 2018 : Hospitalisation. Bruxelles, Belgique : Sciensano ; Numéro de rapport : D/2020/14.044/19. Disponible en ligne : www.enquetesante.be

Drieskens S, **Berete F**, Renard F. Enquête de santé 2018 : Services à domicile et d'aide à domicile. Bruxelles, Belgique : Sciensano ; D/2020/14.440/22. Disponible en ligne : www.enquetesante.be

Van der Heyden J, **Berete F**, Drieskens S. Enquête de santé 2018 : Consommation de médicaments. Bruxelles, Belgique : Sciensano ; D/2020/14/440/25. Disponible en ligne: www.enquetesante.be

Demarest S, Charafeddine R, **Berete F**, Drieskens S. Enquête de santé 2018 : Accessibilité financière aux soins de santé. Bruxelles, Belgique : Sciensano ; Numéro de rapport: D/2020/14.440/23. Disponible en ligne : www.enquetesante.be

Demarest S, Charafeddine R, **Berete F**, Drieskens S. Enquête de santé 2018 : Expérience du patient. Bruxelles, Belgique : Sciensano ; Numéro du rapport : D/2020/14.440/24. Disponible en ligne: www.enquetesante.be

Drieskens S, **Berete F**, Charafeddine R. Accidents. Bruxelles, Belgique : Sciensano ; Numéro de rapport : D/2020/14.440/59. Disponible en ligne : www.enquetesante.be

Braekman E, **Berete F**, Charafeddine R, Drieskens S. Santé sociale. Enquête de santé 2018. Bruxelles, Belgique : Sciensano ; Numéro de rapport : D/2020/14.440/61. Disponible en ligne : www.enquetesante.be

Demarest S, Charafeddine R, **Berete F**, Braekman E, Drieskens S, Gisle L, Hermans L. Cinquième enquête de santé COVID-19. Bruxelles, Belgique ; Numéro de dépôt : D/2020/14.440/96 Disponible en ligne : https://doi.org/10.25608/xcxd-7784

Demarest S, **Berete F**, Charafeddine R, Van der Heyden J. Enquête de santé 2018 : Absence au travail. Bruxelles, Belgique : Sciensano. Numéro de rapport : D/2019/14.440/32. Disponible en ligne : www.enquetesante.be

Charafeddine R, Van der Heyden J, **Berete F**, Demarest S. Enquête de santé 2018 : Connaissances et comportements face au VIH/sida. Bruxelles, Belgique : Sciensano. Numéro de rapport : D/2019/14.440.73. Disponible en ligne : www.enquetesante.be

Charafeddine R, Demarest S, **Berete F**. Enquête de santé 2018 : Littératie en santé. Bruxelles, Belgique : Sciensano. Numéro de rapport : D/2019/14.440.72. Disponible en ligne : www.enquetesante.be

Berete F, Demarest S, Tafforeau J. Enquête de santé 2018 : Dépistage du cancer. Bruxelles, Belgique : Sciensano. Numéro de rapport : D/2019/14.440/75. Disponible en ligne : www.enquetesante.be

Berete F, Tafforeau J, Drieskens S, Demarest S. Enquête de santé 2018 : Vaccination. Bruxelles, Belgique : Sciensano. Numéro de rapport : D/2019/14.440/76. Disponible en ligne : www.enquetesante.be

Demarest S, **Berete F**. Enquête de santé 2018 : Dépistage des facteurs de risque cardiovasculaire et du diabète. Bruxelles, Belgique : Sciensano. D/2019/14.440/74 : Disponible en ligne : www.enquetesante.be

International scientific congresses (oral and poster presentations)

Berete F, Demarest S, Charafeddine R, Meeus P, Bruyère O, Van der Heyden J. Does health literacy mediate the relationship between socioeconomic status and health outcomes? Poster presentation at 16th European Public Health Conference, Dublin, Ireland, 2023.

Berete F, Demarest S, Van Oyen H, Charafeddine R, Bruyère O, Van der Heyden J. Effectiveness of protective measures on dental care utilization: analysis from linked database.

Oral presentation at 16th World Congress on Public Health, Rome, Italy, 2020. (Online event).

Berete F, Demarest S, Charafeddine R, Tafforeau J, Van Oyen H, Bruyère O, Van der Heyden J. Predictors of nursing-home entry for elders in Belgium. Poster presentation at 12th European Public Health Conference, Marseille, France, 2019.

Berete F, Van der Heyden J, Demarest S, Tafforeau J, Van Oyen H, Bruyère O, Renard F. Assessing the validity of self-reported breast cancer screening coverage in the Belgian health interview survey. Poster presentation at European congress of epidemiology, Lyon, France, 2018.