Arch Public Health
2000, 58, 281-294

Evaluation of the encryption procedure and record linkage in the Belgian national cancer registry

by

Van Eycken E. [1], Haustermans K. [2], Buntinx F. [3],
Ceuppens A. [4], Weyler J. [5], Wauters E. [6],
Van Oyen H. [7], De Schaever M. [8],
Van den Berge D. [9], Haelterman M. [1]

Abstract

Aim: *The purpose of this study is to evaluate the encryption and record linkage procedure implemented by the Belgian National Cancer Registry (NCR).*

Corresponding author: Dr. E. Van Eycken, Nationaal Kankerregister, Koningsstraat 217, 1210 Brussels, Belgium, Telephone: 32 2 225 83 93, Fax: 32 2 225 83 97, E-mail: Elizabeth.Vaneycken@kankerregister.org

[1] National Cancer Registry, Koningsstraat 217, 1210 Brussels, Belgium.

[2] Radiotherapy department, University hospital K.U.Leuven, Herestraat 49, 3000 Leuven, Belgium.

[3] Department of General Practice, K.U.Leuven, Kapucijnenvoer 33, 3000 Leuven, Belgium.

[4] National Union of Independent Mutualities, St. Huibrechtsstraat 19, 1150 Brussels, Belgium.

[5] Department of Epidemiology and Community Medicine, University of Antwerp, Universiteitsplein 1, 2610 Antwerp, Belgium.

[6] National Alliance of Christian Mutualities, Haachtsesteenweg 579, 1031 Brussels, Belgium.

[7] Scientific Institute of Public Health, Unit of Epidemiology, J. Wytsmanstraat 14, 1050 Brussels, Belgium.

[8] Radiotherapy department, University hospital R.U.Gent, De Pintelaan 185, 9000 Ghent, Belgium.

[9] Radiotherapy department, Academic Hospital Free University Brussels, Laarbeeklaan 101, 1090 Brussels, Belgium.

Methods: *In order to conform to the privacy legislation, an encryption procedure has been developed for the exchange of cancer notifications between data sources and the NCR. This procedure consists of two steps. 1) A hashing algorithm irreversibly transforms identification data at source into a unique pseudonym. 2) A reversible DES-encryption of the pseudonym is performed at the NCR.*

Record linkage according to the pseudonym has been evaluated with 43.990 records from 16 sources. False negative and false positive matches were estimated using a deterministic linkage with a concatenation variable (month and year of birth, sex, zipcode, initial of first name).

Results: *Linkage based on the pseudonym resulted in 8.936 duplicate registrations. Concatenation variable linkage detected 580 more matches. Additional use of the day of birth in the variable with visual inspection of these 580 probable matches lead to 398 false negative matches (4,3%) due to spelling mistakes in the identification data. Only 4 linked pairs of tumour records were false positive matches (0,01%) due to very common last names.*

Conclusions: *The encryption procedure is feasible. Due to errors in personal data, record linkage based on the pseudonym leads to missed duplicates and an overestimation of cancer incidence. Additional linkage with a concatenation variable including the full date of birth reduces this error, but can only be used as a temporary solution. These results should be taken into account by the authorities to consider a specific law on cancer registration.*

Keywords

Cancer registration, encryption, record linkage.

Introduction

Since 1983, the National Cancer Registry (NCR) centralizes and manages the cancer registry data of Belgian residents. Until 1995 this population-based registry relied on anonymous cancer data from six Belgian health insurance companies. As health insurance is mandatory, more than 99% of the population living in Belgium subscribes to one of these institutions. The health insurance companies collect information on

hospitalized cancer patients from the treating physician on a cancer registration form. Some companies already started cancer registration in the early fourties. At that time cancer was considered a social disease with its specific criteria for reimbursement.

The set up of a cancer registry network with electronic data interchange was started in 1995 to improve the completeness as well as the accuracy of the data. An encryption procedure had to be defined by the NCR in order to conform to the Belgian privacy legislation. This law of December 8th, 1992 states that storage and transmission of an individual's medical data together with identification data is not allowed unless there is informed consent of the patient, or if the data cannot be related to an individual (1). Based on encrypted information, the NCR must be able to recognize multiple notifications of the same individual. A reliable record linkage is very important to avoid over- or under-registration (of disease) due to linkage errors (2). This procedure must also assure the possibility of adding follow-up information on the patients (e.g. date of death).

The purpose of this study was to test the encryption procedure and to evaluate record linkage. We describe the encryption methods used to conform to the Belgian legislation and to the scientific requirements of a cancer registry. The reliability of the NCR record linkage will be estimated by means of performance parameters.

Methods

Network

A cancer registry network was set up to improve the completeness and the validity of the Belgian cancer registry data. The existing registration system based on the participation of the health insurance companies has been extended to other sources since the registration year 1996.

All 7 health insurance companies participated in this study as well as the cancer registry of the province of Antwerp (A.K.R.), the pathology cancer registry of the province of Limburg (LIKAR), the Flemish lung cancer registry (VRGT), 3 hospital registries of university oncology departments (Leuven, Brussels and Ghent) and 3 pathology laboratories (University of Leuven, Sint-Jan hospital of Bruges and Maria-Middelares hospital of Sint-Niklaas). All participants delivered an electronic dataset following the encryption procedure and the minimal dataset requests predetermined by the National Cancer Registry.

Minimal dataset

Each tumour record consists of the information fields set out below. Some fields are obligatory because they are required for a correct interpretation of a unique patient and tumour. These fields are underlined.

Personal data:
- Unique identifier
- Soundex* of first name
- Year of birth
- Sex
- Zipcode of residence
- Date of death

Medical data:
- Date of incidence
- Basis for the diagnosis
- Primary site of tumour and laterality
- Histology and behaviour
- Grade of differentiation
- Clinical and pathological TNM tumour staging
- WHO score at diagnosis (performance status)
- Other staging systems (Dukes, Figo, Ann Arbor, …)
- Treatment(s) within 6 months after diagnosis

Encryption procedure

The encryption protocol consists of 2 steps.

Step 1. A temporary transmission pseudonym (hash code) is generated *at the data source*: 3 identification data (date of birth, sex and last name) are irreversibly transformed using the RIPE-MD 160 hashing algorithm (3). This algorithm is a cryptographic hash function with a 160-bit hash result. The hash code doesn't reveal any information of the input string (personal data) and uncovering is not possible. It is assumed by the developers that this hash function guarantees a sufficient protection for a period of 30 to 50 years (3). All participating sources received the encryption algorithm from the NCR. The source itself is responsible for storing the identification data and the resulting hash code together in

* Phonetic coding system.

a protected file, thereby providing the possibility of exchanging information with the NCR by means of the pseudonym. The input data are extensively checked and converted before starting the hash function. The 3 identification parameters are checked for syntax and format, e.g. all the acceptable characters for the last name were defined beforehand. A semantic control is also included for sex and date of birth: the date must be valid and sex can only be male or female. A conversion program proceeds after this quality control to obtain a maximum of uniformity of the input string; e.g. spaces and two specific punctuation marks are eliminated and lowercases are changed into uppercases.

Step 2. The intermediate transmission pseudonym undergoes a second but reversible encryption *at the NCR*. A 168-bit Triple-DES (data encryption standard)-algorithm with private key at the NCR is being used to prevent dictionary attacks: i.e. someone with access to the database of the NCR applying the hash function on a large file of identities and comparing the resulting codes with the codes stored at the NCR to check whether a certain individual has cancer. This private-key-only method uses the same secret key to encrypt and later decrypt the message. It is therefore also called a symmetric encryption. “Triple” – Des means that the plaintext is encrypted three times, each with a different key.

Record linkage

Record linkage is the matching of multiple records relating to the same person. The record linkage procedure developed by the NCR is based on a deterministic approach using the hash code as a unique identifier for a person. This method generates links based on the agreements (matches) among the hash codes of different files. The hash code, as mentioned above, is the result of an irreversible transformation of the date of birth, sex and last name. It was chosen for the record linkage because it was assumed at the start of the project that this hash code would be a robust identifier. A second reason to choose the hash code as the unique identifier was the prohibition to use personal data due to the stringent privacy law.

After record linkage, predetermined rules were applied to select the most valid information from the matched records to create a composite cancer registration record.

For quality assurance, one must consider two types of linkage errors: homonym and synonym errors (4). Individuals that are erroneously linked or false positive matches lead to homonym errors; failure to link

multiple notifications on the same patient are called synonym errors or false negative matches. Homonym errors will result in a decrease of the estimated incidence of a disease while synonym errors lead to an increase of the incidence. Homonym and synonym error rates are used as performance parameters of the record linkage.

A deterministic method to estimate the homonym and synonym error of the linkage procedure based on the pseudonym was developed at the NCR. Two concatenation variables, consisting of several information fields, were defined. The concatenation variable 1 was made with the initial of first name, month and year of birth, sex and zipcode of residence. A second concatenation variable was defined by the initial of first name, full date of birth, sex and zipcode of residence. Both variables were necessary because one source did not have the full date of birth. The use of 2 consecutive concatenation variables also permitted us to check whether errors could be detected in the exact day of birth.

Tumour or case counts of the computerized record linkage based on the hash codes (test procedure) have been compared with the linkage based on the concatenation variable 2 with visual inspection which is considered as the gold standard. The error rates are calculated for the linkage procedure based on the pseudonym.

Results

A total number of 45.351 tumour records coming from 16 sources was received between July 1999 and February 2000 for the year of incidence 1996. Table 1 gives an overview of the total number of tumour records per source.

TABLE 1
Overview of the total number of tumour records per source

Source	Number of tumour records
Health Insurance Companies (7)	25.989
Provincial Cancer Registry of Antwerp	4.850
Provincial Cancer Registry of Limburg	3.254
Oncology departments of university hospitals (3)	2.670
Pathology laboratories (3)	7.397
Flemish Lung Cancer Registry	1.191
TOTAL	**45.351**
TOTAL MINUS BENIGN AND NON-INCIDENT CASES	**43.990**

After quality control and removal of benign and non-incident cases, 43.990 tumour records were used for the study.

Homonym error rate (figure 1a)

The linkage procedure based on the pseudonym was first applied on the 43.990 tumour notifications. It resulted in 35.054 incident cases and 8.936 duplicate records. The latter must be excluded for case counts when reporting cancer incidence (see table 2). The tumour related information for duplicate records was summarized by a rule-based automated procedure which takes into account the consistency of the notification and the presumed validity of the reporting sources. Manual interventions were undertaken for more complex situations.

False positive matches result from erroneous links of notifications from two different patients and lead to undercounts of tumours because they are erroneously excluded for reporting. To estimate these false positive matches (homonym error), all the linked cases by means of the pseudonym were checked if their concatenation variable 1 also agreed for these matches. When the variable 1 didn't agree on one or more parts, a visual inspection was carried out on the other medical data fields such as topography and histology of the tumour.

As mentioned in table 2, only 4 pairs of records were erroneously linked, giving a very small homonym error rate of 4/34.660 (0.01%). These errors were detected by differences in zipcode of residence and/or differences in initial of the first name. The medical data for these 4 pairs also showed very different tumour characteristics (see table 3).

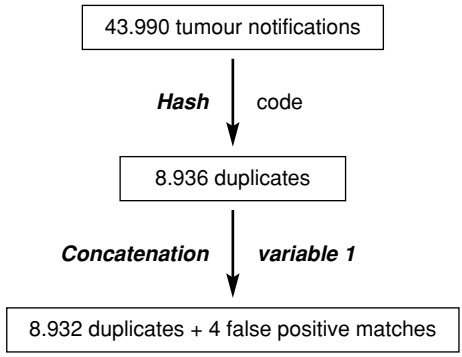

Fig. 1a: Record linkage based on the hash code: precedure to estimate the false positive matches

Both sources involved in these matches communicated with each other by means of the hash code and confirmed that the input parameters (date of birth-sex-last name) for the hashing algorithm were exactly the same (very common last names) for the different patients.

TABLE 2
Tumour counts of the record linkage based on the hash code compared with linkage based on concatenation variable 2 with visual inspection

		Concatenation variable 2 and visual inspection		
		Excluded	Retained	TOTAL
Hash code	Excluded	8.932	4 [1]	8.936
	Retained	398 [2]	34.656	35.054
	TOTAL	9.330 [3]	34.660 [4]	43.990 [5]

[1] Erroneously excluded for tumour counts.
[2] Erroneously retained for tumour counts.
[3] Number of notifications excluded for tumour counts.
[4] Numer of tumours.
[5] Total number of notifications.

TABLE 3
Homonym errors or false positive matches (N = 4), tumour specific information

Source 1	**Source 2**
1. Malignant melanoma of the choroidea	Breast cancer
2. Non Hodgkin lymfoma	Breast cancer
3. Lip cancer	Non Hodgkin lymfoma
4. Endometrium and breast cancer	Ovarian cancer

Synonym error rate (figure 1b)

We then applied the concatenation variable 1 on the same dataset of 43.990 tumour records to detect all possible "missed" links. This variable generated 580 more matches (agreement on concatenation variable 1) than the hash code method. The concatenation variable 1 on its own was not enough to decide whether it was a real match, because too tolerant criteria were used to report a match. Sources were asked to provide the full date of birth (5, 6) for these 580 matches: we proceeded with more strict criteria for a match, adding the day of birth to the variable (concatenation variable 2).

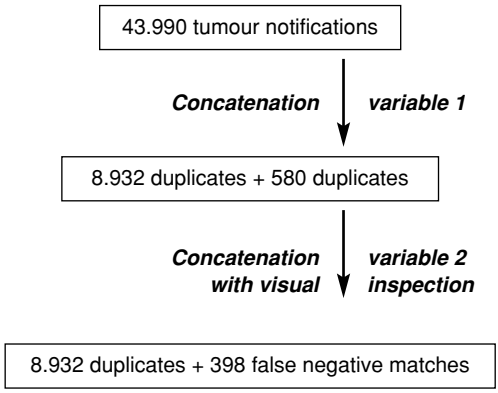

Fig. 1b: Record linkage based on the hash code: precedure to estimate the false negative matches

All the 580 matches were visually inspected using also medical variables such as topography, histology, staging etc. to conclude on a true positive match. It is well known from different studies that manual resolution of a proportion of potential links improves the accuracy of the linkage (7, 8). A total of 398 matches were retained as true positive matches: 28 duplicates were missed due to mistakes in the day of birth, all the other missed matches (370) are explained by spelling or transcription mistakes in the name (e.g. Y versus IJ: 56; K versus CK: 25). This also means that a total of 398 (4,3%) matches out of 9.330 were not detected by the pseudonym procedure and they were erroneously retained for case counts (synonym error rate or false negative matches) which will lead to an overestimation of cancer incidence (+1,1%).

The linkage procedure with the hash code is compared with the results of the concatenation variable 2 in table 2.

Discussion

Stringent confidentiality rules require an encryption procedure for the exchange of tumour notifications between data sources and the NCR. Storage of medical data together with personal data is not allowed unless there is a written consent of the inidividual or if the data cannot be related to an individual. A two steps encryption procedure has been developed with regard to the Belgian privacy legislation and has been approved by the Belgian Privacy Commission and the Belgian National Board of Physicians.

The encryption procedure has to comply with scientific requirements (9). The registry must be able to recognize multiple notifications of the same case. Application of the hash function by the sources on the *same* input parameters will always result in the same hash code. The chance that two distinct patients with *different* input parameters do have the same hash result is negligible ($\sim 10^{-16}$) so that the pseudonym can be used as a unique identifier for one patient (3).

The record linkage procedure should minimize synonym and homonym errors to guarantee the quality of the data. The most striking effects of these errors are over- or underestimation of incidence and survival times: synonym errors lead to an increase of the estimated incidence and survival times while homonym errors will lead to a decrease.

Probabilistic linkage (10, 11) uses conditional probability theory to estimate the likelihood that a pair of records refers to the same individual versus different individuals. It is based on the weights calculated from the comparison of the fields used in the linkage. At this moment, the probabilistic method cannot be applied to the cancer registry because it requires prior knowledge of outcomes in linked and unlinkable pairs (4). In this study, the performance of the record linkage has been evaluated with a deterministic method based on concatenation variables to estimate the error of this hash code linkage. With this direct method, exact values within an individual data field or a group of common fields are matched and record pairs are classified as a link (match) or a non-link (12).

A very small homonym error of 0,01% was found and a 4,3% synonym error rate. The very low rate of homonym error can partly be explained by the fact that the linkage is based on a selected patient population. It is always possible to have two different patients with the same name, date of birth and sex, specifically for common last names. Both results seem fair if we compare them with data from linkage studies in the literature (9, 13, 14, 15, 16). The errors not only depend on the discriminating power of personal identifiers and the quality of the record linkage procedure, but also on the number of cases in the registry, and the number of notifications per case (4). The 4, 3% synonym error rate in this study can be explained by spelling mistakes that were made in the identification data before applying the hash function. High quality input-data are essential to reduce spelling errors. Moreover, certain errors can be avoided when the source applies phonetic rules (e.g. change every CK into K and change IJ into Y) or a phonetic coding system as has been developed in the United States (Soundex) (15). These more restrictive

definitions of case identity will result in a reduction of false negative matches but will lead to an increase of false positive matches.

It is well known from the literature that linkage based on a unique personal identification number, such as is used in Scandinavian countries, is more accurate because it avoids spelling errors (17, 18, 19). It is even better when this personal identification number, such as the social security number or the national registry number, is a machine readable number. The use of such a personal number requires the availability of that number at every source of the cancer network. Until now, the law prevents such an approach in Belgium.

A temporary but feasible and valuable alternative is the collection of the full date of birth for every tumour record. This allows record linkage based on concatenation variable 2 in order to detect and reduce the number of false negative matches.

Both increasing the discriminating power of the identifier(s) for record linkage and reducing the errors in coding or reporting will minimize the errors.

Conclusions

The NCR was able to recognize multiple notifications describing the same individual. Only 4 pairs of tumour records were false positive matches (0.01%). Record linkage based on the hash code leads to missed duplicates due to spelling and transcription errors in identification data. The additional linkage with a concatenation variable including the full date of birth reduces this error, but can only be used as a temporary solution. These results should be taken into account by the authorities to consider a specific law on cancer registration.

Acknowledgments

This work was supported by the Flemish Government and the Flemish Minister of Preventive and Social Health Care.

We thank M. Renson and C. Platteau for their help with the preparation of the manuscript.

Samenvatting

Doelstelling: In deze studie wordt de anonimiseringsprocedure en de gegevenskoppeling van het Belgische Nationaal Kankerregister (NKR) geëvalueerd.

Methoden: Om te voldoen aan de privacy wetgeving, werd een encryptieprocedure ontwikkeld voor het uitwisselen van kankerregistratiegegevens tussen bronnen en het NKR. Eerst worden door de bron een aantal persoonsgegevens onomkeerbaar versleuteld tot een pseudoniem door middel van een hashing algoritme. Vervolgens gebeurt op het NKR een reversiebele DES-encryptie van het pseudoniem.

De gegevenskoppeling gebaseerd op het pseudoniem werd geëvalueerd met 43.990 tumor-informatielijnen van 16 verschillende bronnen. Vals negatieve en vals positieve koppelingen werden opgezocht aan de hand van een deterministische linkageprocedure met een concatenatievariabele (geboortejaar en geboortemaand, geslacht, postcode en initiaal van de voornaam).

Resultaten: De koppeling op basis van het pseudoniem leverde 8.936 dubbele registraties op. Linkage met de concatenatievariabele detecteerde er 580 méér. Toevoeging van de geboortedag aan de concatenatievariabele in combinatie met visuele inspectie van deze 580 mogelijke dubbels, resulteerde in 398 vals negatieve koppelingen (4,3%) ten gevolge van spellingsfouten in de identificatiegegevens van de bron. Slechts 4 gekoppelde tumoren waren verkeerdelijk samengevoegd (0,01%) door zeer frekwent voorkomende familienamen.

Conclusies: De encryptie-procedure van het NKR is werkbaar. Ten gevolge van spellingsfouten in de persoonsgegevens leidt koppeling door middel van het pseudoniem tot gemiste dubbels en een overschatting van het aantal tumoren. Bijkomende linkage met een concatenatievariabele die de volledige geboortedatum bevat, reduceert deze fout, maar kan enkel als een tijdelijke oplossing worden beschouwd. De resultaten van deze studie moeten het beleid aanzetten om een wet op kankerregistratie te overwegen.

Résumé

Objectif: Dans cette étude, on évalue la procédure d'encryptage et le couplage de données du Registre National du Cancer (RNC) belge.

Méthodes: En conformité avec la loi sur la protection de la vie privée, une procédure d'encryptage a été développée pour l'échange des données d'enregistrement du cancer entre les sources et le RNC. La source encrypte irréversiblement les données d'identification en pseudonyme avec un algorithme de hashage. Ensuite, le RNC effectue un encryptage réversible DES du pseudonyme. Le couplage de données basé sur le pseudonyme a été évalué sur 43.990 lignes d'information tumeurs provenant de 16 sources. Des couplages faux négatifs et faux positifs ont été recherchés par une procédure de couplage par concaténation des variables (année de naissance, mois de naissance, sexe, code postal et initiale du prénom).

Résultats: Le couplage basé sur le pseudonyme produit 8.936 enregistrements doubles. Le couplage par concaténation des variables en détecte 580 de plus. L'intégration du jour de naissance dans la concaténation, combinée avec l'inspection visuelle de ces 580 doublons produit 398 couples faux négatifs (4,3%) en raison de fautes d'ortographe dans les données d'identification à la source. Seules 4 tumeurs couplées sont faussement groupées (0,01%) par la présence de noms de famille très fréquents.

Conclusions: La procédure d'encryptage du RNC est faisable. Suite aux fautes d'ortographe, le couplage par le pseudonyme produit des doublons manqués et par conséquent une surestimation du nombre des tumeurs. Un couplage complémentaire par concaténation incluant la date de naissance complète réduit cette erreur. Les résultats de cette étude devraient motiver les autorités à considérer une loi spécifique pour l'enregistrement du cancer.

References

1. Wet ter Bescherming van de Persoonlijke Levenssfeer ten opzichte van de verwerking van persoonsgegevens, 8 december 1992. Belgisch Staatsblad, 1993.
2. BRENNER H, SCHMIDTMANN I. Effects of record linkage errors on disease registration. Methods Inf Med 1998, 37: 69-74.
3. DOBBERTIN H, BOSSELAERS A, PRENEEL B. Ripe MD-160, a strengthened version of RIPEMD. Fast Software encryption 1996, LNCS 1039: 71-82.
4. BRENNER H, SCHMIDTMANN I. Determinants of homonym and synonym rates of record linkage in disease registration. Meth Inform Med 1996, 35: 19-24.
5. VAN DEN BRANDT PA, SCHOUTEN LJ, GOLDBOHM RA, DORANT E, HUNEN PMH. Development of a record linkage protocol for use in the Dutch cancer registry for epidemiological research. Int J Epidemiol 1990, 19: 553-558.
6. ROOS LL, WAJDA A. Record linkage strategies. Part I: Estimating information and evaluating approaches. Methods Inf Med 1991, 30: 117-123.
7. MACLEOD MC, BRAY CA, KENDRICK SW, COBBE SM. Enhancing the power of record linkage involving low quality personal identifiers: use of the best link principle and cause of death prior likelihoods. Comput Biomed Res 1998, 31: 257-270.
8. SIMONATO L, ZAMBON P, RODELLA S et al. A computerised cancer registration network in the Veneto region, North-east of Italy: a pilot study. Br J Cancer 1996, 73: 1436-1439.
9. POMMERENING K, MILLER M, SCHMIDTMANN I, MICHAELIS J. Pseudonyms for Cancer Registries. Meth Inform Med 1996, 35: 112-121.
10. JARO MA. Probabilistic linkage of large public health data files. Stat Med 1995, 14: 491-498.
11. WAJDA A, ROOS LL, LAYEFSKI M, SINGLETON JA. Record linkage strategies: Part II. Portable software and deterministic matching. Methods Inf Med 1991, 30: 210-214.
12. BLACK RJ, SIMONATO L, STORM HH, DEMARET E. Automated data collection in cancer registration. Lyon, IARC technical reports N° 32, 1998, 7-11.
13. LIU S. Development of record linkage of hospital discharge data for the study of neonatal readmission. Chronic Dis Can 1999, 20: 77-81.
14. MUSE AG, MIKL J, SMITH PF. Evaluating the quality of anonymous record linkage using deterministic procedures with the New York State AIDS registry and a hospital discharge file. Stat Med 1995, 14: 499-509.

15. QUANTIN C, BOUZELAT H, ALLAERT F A, BENHAMICHE A M, FAIVRE J, DUSERRE L. Automatic record hash code and linkage for epidemiological follow-up data confidentiality. Meth Inform Med 1998, 37: 271-277.
16. KJAERHEIM K. Occupational cancer research in the Nordic countries. Environ Health Perspect 1999, 107: 233-238.
17. WIKLUND K, EKLUND G. Reliability of record linkage in the Swedish cancer-environment register. Acta Radiol Oncol 1986, 25: 11-14.
18. ROOS LL, WALLD R, WAJDA A, BOND R, HARTFORD K. Record linkage strategies, outpatient procedures, and administrative data. Med Care 1996, 34: 570-582.
19. STORM HH. Completeness of cancer registration in Denmark 1943-1966 and efficacy of record linkage procedures. Int J Epidemiol 1988, 17: 44-49.