Arch Public Health
1998, 56, 345-361

Investigation of the clustering effect in the Belgian Health Interview Survey 1997

by

Renard D. [1], Molenberghs G. [1], Van Oyen H. [2] and Tafforeau J. [2]

Abstract

This paper investigates the effect of clustering in the first Health Interview Survey (HIS) that took place in Belgium in 1997. In this survey 10,221 individuals were interviewed using a stratified multistage clustered sampling procedure. Clustering arises at two levels in the HIS, within municipalities and within households. Its effect and magnitude on some selected continuous and discrete items are studied from a multilevel modeling perspective. This model-based approach fully acknowledges and takes advantage of the hierarchical structure of the data, and is to be contrasted with the more traditional, design-based approach which views the population structure as a nuisance factor. The effect of weighting in this context is also investigated following Pfeffermann et al. (1).

Key-words

Health Interview Survey; clustering; multilevel models; weighting.

Address correspondence to :
[1] Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium.
[2] Scientific Institute for Public Health, Rue J. Wytsman 14, B-1050 Brussels, Belgium.

1. Introduction

In 1997 the first Health Interview Survey (HIS) took place in Belgium, following a stratified multistage clustered sampling procedure. This means, in particular, that selection within each stratum was carried out in several stages, the sampling units at each stage being subsampled from the larger units (or clusters) chosen at the previous stage. As a result of this sampling design, the individuals in the population were included in the sample with differing probabilities. Typically, analyses of data arising from complex sample surveys are adjusted by methods that incorporate sampling weights, usually defined as the reciprocals of the sample inclusion probabilities. These sampling weights effectively represent the number of individuals in the population that each sampled individual represents (2).

Another consequence of the multistage selection, which is not accounted for by the use of weights, is that clusters selected at various stages constitute relatively homogeneous groups. Clustered data ordinarily exhibit intra-class correlation because units from the same cluster tend to be more alike than units from different clusters, thus violating the usual independence assumption underlying many common statistical methodologies. Whatever method of analysis is used, it should address this issue in order to obtain valid statistical inference (e.g., correct variances of the regression coefficients) and special procedures have been developed to deal with clustering in a sample survey framework, for example when comparing mean values or fitting classical regression models (3).

This type of procedure is encompassed in the traditional, design-based philosophy and typically constitutes an *ad hoc* correction to account for the sampling design. In this approach the population structure, insofar as it is mirrored in the sampling design, is seen as a nuisance factor (4). On the contrary, the multilevel modeling approach views the population structure as of potential interest in its own right and intimately embeds this structure in the model itself. In addition, a model-based approach enables one to incorporate design-related information directly into the model, thus obviating the need to carry out special procedures to adjust for the effects of the sampling design.

On the grounds of these considerations, we shall endorse the latter viewpoint to investigate the effect and magnitude of clustering in the HIS. In Section 2 we outline the main design aspects of the HIS. Section 3 reviews multilevel modeling techniques for both continuous and discrete response data and discusses weighting in this context, while the next section is devoted to applying these methods to the HIS. We address some comments and conclusions in the last section.

2. Overview of the HIS Sampling Design

A detailed account of the HIS sampling design has been given elsewhere (5). See Van Oyen *et al.* (6) for a more concise description. In this section, we briefly outline the main aspects of the final sampling scheme for the selection of the households (HHs) and respondents in the HIS.

Basically, the sampling procedure can be seen as a combination of different sampling techniques: stratification, multistage sampling and clustering. Stratification is performed at the regional level (Flemish, Walloon and Brussels regions) and at the provincial level. A further refinement concerns the German community which has been considered a proper entity on its own. This stratification aims at achieving a geographical spread of the interviews and overall, gives rise to 12 strata. The quota of interviews were also evenly distributed over quarters of the study year to obtain reasonable spread over time.

Next, the individuals' sample is selected in three stages within each stratum. The first stage, yielding primary sampling units (PSU), consists of municipalities and sampling is carried out proportionally to (population) size. Whenever a municipality is selected (and it can be more than once), a group of 50 persons is to be interviewed within this municipality. The next stage of random selection operates on HHs (secondary sampling units or SSU) according to a clustered systematic sampling procedure upon ordering of the HHs by statistical sector, size and age of the reference person. At this level, matching HHs are provided in case a HH refuses to participate. Finally, individuals or tertiary sampling units (TSU) are selected within HHs in such a way that 4 persons at most are interviewed in each HH and the reference person and his/her partner are automatically selected.

3. Multilevel Models

Many sets of data collected in human and biological sciences have a hierarchical or clustered structure (a hierarchy consists of units grouped at different levels). Examples of such data structures abound: individuals are grouped into HHs, workers into workplaces, animals into litters, or subjects can be studied repeatedly over time, thus yielding measurements grouped within individuals. Obviously, one can extend such elementary structures to any number of levels.

The existence of hierarchies in some data structures is neither accidental nor ignorable and it has long been recognized that the grouping or clustering induced by such hierarchies presents particular problems due to the lack of independence between observations. In the 1980s a number of researchers have introduced systematic approaches to the statistical analysis of hierarchically nested data. This has resulted, in the early 1990s, in a core set of well-developed, established techniques and widely available software packages that could be applied to the fitting of multilevel models in an efficient manner. This is especially true for responses that are continuously distributed. In the last decade much effort has been initiated to cope with discrete response data in models involving random effects, and this has led to the development of generalized linear mixed models which is still a field of active research.

The remainder of this section is intended to give a short overview of multilevel modeling. We also discuss weighting in this context. For more comprehensive accounts on the subject, see Bryck and Raudenbush (7), Longford (8) and Goldstein (4). A more recent introduction is given by Kreft and de Leeuw (9).

3.1 The Multilevel Linear Model

This section focuses on multilevel models for outcome variables that are continuously distributed. We consider the case of a three-level model with application to the HIS data in mind. Thus, suppose we have measurements Y_{ijk} on a continuous response variable for the ith individual from the jth HH in the kth municipality.

A multilevel model contains, in general, variables measured or defined at different levels of the hierarchy and allows regression coefficients to be random at any of these levels. As such, it is a special case of the general linear mixed model (10, 11) and hence can be written as

$$Y = X\beta + Zu + \varepsilon, \tag{3.1}$$

where Y is the vector of responses, X is a model matrix for the fixed effects with corresponding parameter vector β, Z is a model matrix for the random effects u, and ε is a vector of error terms. The vector u of random effects and the error term ε are assumed to be mutually independent. The variance matrix of ε is ordinarily assumed to be (a multiple of) $\sigma^2_e I_N$ with N the total number of observations, but does not need to be so. When the top level units are independent, it follows that the variance matrix V = Var (Y) has a patterned (block diagonal) structure which can be exploited in the calculations (12).

Technical Note

With the further distributional assumption of normality for u and ε, parameter estimation can proceed by maximization of the likelihood function. A diversity of algorithms have been proposed which comprise iterative generalized least squares (12), Fisher scoring (13) or Expectation-Maximization (14). Alternatively, one may employ the method of generalized estimating equations (GEE) introduced by Liang and Zeger (15), which is a generalization of quasi-likelihood estimation (16) and focuses primarily on modeling the mean structure rather than exploring the random component of the model. This procedure is, however, most useful in the generalized linear model framework. More recently, the treatment from a fully Bayesian perspective has become computationally feasible with the development of Markov Chain Monte Carlo (MCMC) methods, especially Gibbs sampling (17).

3.2 Multilevel Models for Discrete Response Data

We shall restrict discussion to the case of binary (or, more generally, binomial) responses. Following Rodríguez and Goldman (18), the multilevel logit model is obtained by assuming that, conditional on a vector of random effects u, the elements of Y are independent Bernoulli random variables with probabilities $\mu_{ijk} = P[Y_{ijk} = 1|u]$ satisfying

$$\eta = \text{logit}\ (\mu) = X\beta + Zu, \tag{3.2}$$

where u is assumed to be normally distributed. Note that (3.2) may equally be derived from a multilevel linear model using a latent variable formulation.

The conditional likelihood function assumes the usual binomial form, where conditioning is done on the random effects. The major challenge with model (3.2) is to obtain the unconditional likelihood since we need to integrate out the random effects, which unfortunately yields an intractable expression. Numerical integration can be accomplished for relatively simple models but soon becomes computationally intensive as complexity (dimension of the random effects) grows. Different procedures have been proposed to circumvent the problem and most of them rely on approximations.

Technical Note

Goldstein (19) for instance writes his model as

$$Y = \mu(\eta) + \varepsilon, \tag{3.3}$$

where μ satisfies (3.2) and the mean-zero error term ε represents the element level variability and is defined so as to conform to the assumption $Y_{ijk} \sim$ Binomial $(1, \mu_{ijk})$. He proposes to linearize the mean function $\mu(\eta)$ using a Taylor expansion around the fixed part predictor (i.e., setting $u = o$). Upon regrouping the procedure entails fitting multilevel linear models repeatedly. This procedure has also been referred to as marginal quasi-likelihood (MQL) by Breslow and Clayton (20). Rodríguez and Goldman (18) point out that MQL can be seriously biased, in response to what Goldstein and Rasbash (21) improve the procedure by adding estimated residuals to the fixed part predictor in the Taylor expansion – this amounts to the penalized quasi-likelihood (PQL) procedure of Breslow and Clayton (20) – and considering second-order terms (PQL2). A notable feature of MQL and PQL procedures is that they can easily be extended to accomodate extra-binomial variation, in which case they yield results equivalent to the pseudo-likelihood approach of Wolfinger and O' Connell (22).

3.3 Weighting in Multilevel Models

The issue of weighting in multilevel models has not been extensively investigated in the literature until quite recently (see Graubard and Korn (2) and Goldstein (4) for restricted discussions, and Pfeffermann *et al.* (1) for a more thorough treatment in the case of linear models). A reason might be that sampling schemes are commonly ignored in multilevel analyses of survey data since multilevel models enable one to incorporate certain characteristics of the sampling design as covariates, such as strata indicators or size variables. This argument breaks down when the relevant information is not made available to the analyst or when it is not scientifically meaningful to be included in the model.

It should be emphasized that weighting in multilevel models is not a trivial extension of conventional methods of weighting (1). One key feature of the multilevel approach is that sample inclusion probabilities can be defined at any stage of the hierarchy, conditionally on the membership to higher clusters. Thus, municipality κ is selected with inclusion probability π_κ, HH j is selected with probability $\pi_{j|\kappa}$ within municipality κ, and individual i is sampled with probability $\pi_{j|\kappa}$ within HH j selected from municipality κ. Unconditional selection probabilities can be derived from appropriate products of conditional probabilities (e.g., $\pi_{j\kappa} = \pi_\kappa \pi_{j|\kappa}$ denotes the probability that municipality κ is sampled and that, within this municipality, HH j is selected). These selection probabilities, in turn, permit to define corresponding sample weights by taking reciprocals.

Technical Note

The approach Pfeffermann *et al.* advocate consists of replacing in the sample estimators each sum over units from a given level by a correspondingly weighted sum, using conditional weights defined at the same level. They achieve this weighting by considering a two-step procedure of which the first step merely entails a transformation of the data (more specifically, of the covariates specifying the random part of the model). It turns out that this simple step alone almost achieves the sought-after goal, which consequently makes it very attractive insofar as standard software packages can be employed for its implementation. Pfeffermann *et al.* find, on the grounds of limited simulations, the corresponding estimator to perform fairly well, except when the sampling mechanism is informative (i.e., when the sample inclusion probabilities are related to the error terms and hence to the response data) and in some other special instances. For the sake of simplicity, we consider solely this restricted form of the weighting procedure in the sequel.

4. Application to the HIS

This section aims at applying the multilevel modeling methodology to selected items from the HIS. For modeling purposes we need to choose a set of explanatory variables to be included in each of our models. These are region, sex, age (eight categories), education (five categories) and HH income (five categories).

For the computational aspects of this work, we used the stand-alone software package MLwiN (23). This package has been specifically designed for the fitting of multilevel models and can deal both with continuous and discrete response data. For the latter, we also considered the program MIXOR (24).

4.1 Continuous Response Data

The following items were considered: body mass index (BMI) and VOEG score. The VOEG score is derived from a 23-item questionnaire which produces an inventory of various complaints. The items are 0/1 coded and summed up to obtain the final score (higher scores are indicator of more pronounced morbidity). This score is, strictly speaking, an ordered categorical variable. However, the number of categories being large, it was treated as a continuous covariate. Log transformations, specifically ln(BMI) and ln(VOEG+1), were applied to these variables to normalize their distribution. Overall, 9118 out of 10,221 (89%) and 7570 out of 8560 (88%) observations were available on the selected covariates and on BMI and VOEG score, respectively. Notice that the VOEG questionnaire was addressed solely to people aged above 15 years.

The following model was fitted to these data:

$$Y_{ijk} = x'_{ijk}\beta + v_k + u_{jk} + \varepsilon_{ijk}, \tag{4.4}$$

where v_k, u_{jk} and ε_{ijk} are variance components defined respectively at the municipality, the HH and the individual level. Unlike the vector notation in (3.1), we make use of a scalar notation here. Formaly, (4.4) can be rewritten as (3.1) by stacking the responses Y_{ijk} and the row vectors of fixed-effects x'_{ijk} to form Y and X respectively. Likewise, the vector u of random effects can be obtained by stacking the municipality and HH effects (v_k and u_{jk}) and the Z matrix by defining indicators of municipalities and HHs to conform with the structure of u.

The interpretation of the above variance components can be done in terms of intra-class correlations (ICCs). If we suppose that $v_k \sim N(0,\sigma^2_v)$, $u_{jk} \sim N(0,\sigma^2_u)$ and $\varepsilon_{ijk} \sim N(0,\sigma^2_e)$ and that these three components are independent of each other, then the intra-municipality correlation is defined as

$$\rho_{MUN} = \frac{\sigma^2_v}{\sigma^2_v + \sigma^2_u + \sigma^2_e}, \tag{4.5}$$

while the intra-HH correlation is equal to

$$\rho_{HH} = \frac{\sigma^2_v + \sigma^2_u}{\sigma^2_v + \sigma^2_u + \sigma^2_e}, \tag{4.6}$$

Thus, the intra-class correlation at a given level is the proportion of total variability in the outcome variable that lies between the units at that level, and hence is a measure of (within) group homogeneity (the more homogeneous the groups, the higher the intra-class correlation).

Besides this measure of clustering, there is also some interest in comparing performance of weighted and unweighted estimators and assessing to what extent inclusion of the design-related information into the model might influence their performance. In particular, one can examine whether the effect of weighting is diminished when variables determining the sampling rates are included as covariates.

Before presenting the results, we first elaborate on the design information we considered for inclusion into the model. As outlined in Section 2, we can first include strata indicators for provinces (hereby referring to the usual 10 Belgian provinces plus Brussels and the German community) and quarters. We can then supplement these indicators with variables which characterize sampling at each of the levels. Thus, within a province of P individuals, the population size S_κ of municipality κ along with the number g of groups of 50 persons to be interviewed determine the selection probability of that municipality ($\pi_\kappa = gS_\kappa/P$) and can therefore be considered as potential covariates. Likewise, HH size and status of HH members (specifically, whether a HH member is the reference person or his/her partner) enable one to calculate selection probabilities at the individual level and can be included in a model. For simplicity, we have not considered sampling information at the HH level. It should be noticed, however, that there were no differential selection probabilities at that level.

Tables 1 and 2 summarize the results. In each case, the first part of the table refers to the model containing the explanatory variables of interest, but adjusted for the design variables described above (note that parameter estimates of these variables are not reported for reasons of readibility), whereas the second part concerns the model with explanatory variables of interest only. In each case, weighted and unweighted estimators were computed, where in the former (conditional and unconditional) weights at each level were rescaled using the corresponding weight averages. The empirical or sandwich estimators (4, 15) are reported for the standard errors of the parameter estimates because the appropriate covariance matrix cannot be computed directly in MLwiN. Observe that the regional effects could not be estimated when including design variables due to the presence of provinces in the model, which form a finer partition of regions. Instead, we report for each region the average provincial effect within this region.

TABLE 1

Parameter estimates for the ln (BMI) data†

Parameter	Including design variables: Unweighted Estimator	Including design variables: Weighted Estimator	Not including design variables: Unweighted Estimator	Not including design variables: Weighted Estimator
β Intercept	2.8249 (0.0147)	2.8281 (0.0220)	2.8515 (0.0125)	2.8507 (0.0208)
Sex (Male)	0.0359 (0.0033)	0.0409 (0.0048)	0.0356 (0.0033)	0.0397 (0.0049)
Age (reference: 0-14)				
15-24	0.2245 (0.0068)	0.2481 (0.0130)	0.2510 (0.0068)	0.2528 (0.0140)
25-34	0.3182 (0.0087)	0.3299 (0.0123)	0.3458 (0.0073)	0.3555 (0.0086)
35-44	0.3472 (0.0088)	0.3648 (0.0141)	0.3810 (0.0074)	0.3981 (0.0094)
45-54	0.3829 (0.0096)	0.3871 (0.0130)	0.4164 (0.0071)	0.4176 (0.0110)
55-64	0.4086 (0.0101)	0.4268 (0.0147)	0.4391 (0.0072)	0.4563 (0.0111)
65-74	0.3995 (0.0102)	0.3990 (0.0149)	0.4284 (0.0087)	0.4273 (0.0191)
75+	0.3404 (0.0118)	0.3458 (0.0159)	0.3671 (0.0088)	0.3713 (0.0159)
Income (reference: < 20,000)				
20,000-30,000	0.0007 (0.0074)	– 0.0310 (0.0159)	– 0.0018 (0.0073)	– 0.0316 (0.0158)
30,000-40,000	0.0079 (0.0066)	– 0.0160 (0.0138)	– 0.0040 (0.0065)	– 0.0177 (0.0132)
40,000-60,000	0.0022 (0.0071)	– 0.0346 (0.0187)	– 0.0026 (0.0069)	– 0.0369 (0.0158)
> 60,000	– 0.0004 (0.0076)	– 0.0281 (0.0200)	– 0.0077 (0.0075)	– 0.0338 (0.0201)
Education (reference: No diploma)				
Primary	– 0.0303 (0.0101)	– 0.0195 (0.0146)	– 0.0365 (0.0100)	– 0.0251 (0.0148)
Lower secondary	– 0.0487 (0.0107)	– 0.0366 (0.0158)	– 0.0523 (0.0101)	– 0.0390 (0.0150)
Higher secondary	– 0.0630 (0.0150)	– 0.0472 (0.0169)	– 0.0645 (0.0099)	– 0.0454 (0.0157)
Higher	– 0.0911 (0.0104)	– 0.0789 (0.0160)	– 0.0914 (0.0100)	– 0.0763 (0.0153)
Region (reference: Brussels)				
Flanders	0.0075††	0.0133††	0.0054 (0.0061)	0.0101 (0.0083)
Wallonia	0.0195††	0.0201††	0.0224 (0.0059)	0.0218 (0.0087)
σ^2_v	0.0001 (0.0001)	0.0001 (0.0001)	0.0001 (0.0001)	0.0003 (0.0002)
σ^2_u	0.0039 (0.0005)	0.0040 (0.0010)	0.0038 (0.0005)	0.0040 (0.0011)
σ^2_e	0.0217 (0.0011)	0.0211 (0.0042)	0.0218 (0.0011)	0.0213 (0.0042)
ρ_{HH}†††	0.154 (0.019)	0.165 (0.033)	0.153 (0.019)	0.167 (0.034)

† Empirical standard errors are given in parentheses.
†† Average over the provincial effects (see main text).
††† Standard errors were calculated using the delta method.

We see that there is generally good agreement (within 95% confidence limits) between weighted and unweighted estimators, but that standard errors of the weighted estimators are subject to a substantial (sometimes more than twofold) loss of efficiency. This conclusion can be drawn whether or not we include some design-related information into the model. Of course, some loss of efficiency is to be expected, as it can be in general for any weighted estimator. Korn and Graubard (25) for instance illustrate the latter point in simple situations and clearly point out that the variance of the weighted estimator becomes larger as the

TABLE 2
Parameter estimates for the ln (VOEG + 1) data†

Parameter	Including design variables: Unweighted Estimator		Including design variables: Weighted Estimator		Not including design variables: Unweighted Estimator		Not including design variables: Weighted Estimator	
β Intercept	1.867	(0.077)	1.670	(0.135)	1.770	(0.064)	1.587	(0.117)
Sex (Male)	− 0.244	(0.018)	− 0.219	(0.029)	− 0.246	(0.018)	− 0.224	(0.029)
Age (reference: 14-25)								
25-34	0.027	(0.042)	0.034	(0.054)	0.069	(0.032)	0.126	(0.047)
35-44	0.173	(0.042)	0.223	(0.066)	0.212	(0.032)	0.331	(0.058)
45-54	0.185	(0.048)	0.147	(0.061)	0.238	(0.038)	0.269	(0.053)
55-64	0.248	(0.047)	0.245	(0.062)	0.313	(0.038)	0.389	(0.051)
65-74	0.171	(0.046)	0.192	(0.067)	0.245	(0.037)	0.340	(0.050)
75+	0.195	(0.050)	0.248	(0.085)	0.272	(0.043)	0.396	(0.070)
Income (reference: < 20,000)								
20,000-30,000	0.124	(0.044)	0.194	(0.106)	0.133	(0.045)	0.207	(0.110)
30,000-40,000	0.051	(0.045)	0.074	(0.114)	0.064	(0.046)	0.095	(0.120)
40,000-60,000	0.039	(0.043)	0.091	(0.110)	0.049	(0.044)	0.111	(0.118)
> 60,000	− 0.073	(0.050)	0.014	(0.115)	− 0.059	(0.051)	0.059	(0.117)
Education (reference: No diploma)								
Primary	− 0.058	(0.060)	− 0.001	(0.070)	− 0.046	(0.060)	− 0.006	(0.068)
Lower secondary	− 0.014	(0.060)	0.006	(0.072)	0.002	(0.060)	− 0.002	(0.071)
Higher secondary	− 0.085	(0.060)	− 0.001	(0.068)	− 0.067	(0.061)	0.002	(0.067)
Higher	− 0.140	(0.055)	− 0.050	(0.075)	− 0.125	(0.056)	− 0.050	(0.077)
Region (reference: Brussels)								
Flanders	− 0.224††		− 0.234††		− 0.290	(0.0061)	− 0.302	(0.047)
Wallonia	− 0.051††		− 0.060††		− 0.016	(0.0059)	− 0.026	(0.050)
σ^2_v	0.013	(0.004)	0.013	(0.005)	0.019	(0.005)	0.021	(0.006)
σ^2_u	0.140	(0.015)	0.123	(0.041)	0.142	(0.015)	0.127	(0.040)
σ^2_e	0.446	(0.019)	0.461	(0.093)	0.446	(0.019)	0.462	(0.092)
ρ_{HH}†††	0.256	(0.024)	0.228	(0.060)	0.266	(0.024)	0.242	(0.055)

† Empirical standard errors are given in parentheses.
†† Average over the provincial effects (see main text).
††† Standard errors were calculated using the delta method.

sample weights exhibit more variability. In the HIS, (scaled) weights ranged from about 0.02 to about 10 at each level, thus revealing much variability. Should we consider the two-level model ignoring municipalities, weights would exhibit a greatly reduced variability (actually, none at the HH level and very little at the individual level) and would therefore hardly affect the weighted estimators (results not shown). Interestingly, the aforementioned increase in standard errors was not so drastic for the default model-based estimators provided by MLwiN. We have not reported these estimates, however, because their use is misleading.

The variance components estimates are very similar for the BMI data and a bit more discrepant for the VOEG score. Both show that there is little intra-municipality correlation, whereas at the HH level, ICC is moderate (ρ_{HH} is about 0.16 for BMI and 0.24 for VOEG score). Standard errors associated with the intra-HH correlation parameter were obtained using the delta method for a ratio of two parameters (26).

4.2 Binary Response Data

In this section, we elaborate on a couple of questions raised by the discreteness of the data and then present the results on two items from the HIS.

In Section 3.3, we have discussed weighting in the multilevel linear model, assuming implicitly that the response variable were normally distributed. One might argue that an analogous procedure applies to the multilevel logit model (MQL or PQL estimation procedures entail iterated fitting of linear models) with the following exception to the weighting rule. Since the element level defines the binomial variation, a method of incorporating the weights at this level is to multiply the denominator (i.e., the “number of trials” in binomial terms) by the element level weights. Weighting at higher levels remains unchanged. Because this procedure has not been extensively investigated, we do not consider weighting further here. As a consequence, it is no longer deemed necessary to maintain a municipality-level component in our model, for there is little indication of variability at this level. Accordingly, in the remainder of this section we restrict attention to a two-level model wherein individuals are nested within households.

A second difficulty raised by the model formulation (3.2)-(3.3) is to carry over the concept of ICC. Indeed, the fact that the element level variance component is on the probability scale, whereas the higher level components are on the logit scale, precludes direct comparison of variance components in such models and hence the use of formulae similar to (4.5) or (4.6). Recall, however, that (3.2) can equivalently be obtained from a multilevel linear model using a latent variable formulation. Hedeker and Gibbons (27) consider this scenario and utilize a normal or logistic distribution for the residual error term. When the logistic model is assumed, the residual variability is equal to $\pi^2/3$ and gives rise to the following expression for the intra-HH correlation:

$$\rho_{HH} = \frac{\sigma^2_u}{\sigma^2_u + \pi^2/3}, \tag{4.7}$$

where σ^2_u is the variance associated with the HH random component. Thus, to consider ICC in multilevel binary response models, one must make reference to the underlying continuous (unobserved) variable that generates the dichotomous outcome. In some cases, this is a reasonable assumption, but in others it may not be. Hedeker and Gibbons' methodology has been implemented in a program called MIXOR which, albeit limited to 2-level models, accomplishes numerical integration to calculate the likelihood.

For comparison purposes, we further consider the GEE approach of Prentice (28). This method emphasizes the modeling of the mean structure (e.g., logit $(\pi_{ij}) = x'_j\beta$ for the ith individual from the jth HH) and accounts for the intra-HH correlation by means of a working correlation matrix, unlike the preceding approach which stipulates the covariance structure through a HH-specific (random) effect. We assume an exchangeable correlation structure (i.e., observations are assumed to be equally correlated, with correlation ρ). The interpretation of this parameter can be done in terms of intra-HH correlation.

We can now concentrate on the application of these modeling strategies to the HIS. We consider two items, namely, whether or not a person has a steady general practitioner (GP), and the so-called subjective or perceived health, wich consists of a self-assessment of an individual's health status. The latter was dichotomized as good to very good versus other. There were 9328 out of 10,221 (91%) and 7281 out of 8560 (88%) observations available on both the explanatory variables and steady GP and subjective health, respectively (notice that solely people above 15 were enquired about their subjective health).

Table 3 displays the results for both items. Columns denoted GEE1.5 refer to the GEE approach of Prentice (28), while GLMM refers to the generalized linear mixed model fitted by MIXOR. Two things are apparent in this table: the shrinkage of GEE1.5 regression coefficients towards zero compared to GLMM, and the somewhat dissimilar estimates of intra-HH correlation. The former is a well-known phenomenon that discriminates between marginal and random-effects models (29). These authors show that the discrepancy increases with the size of ICC, as can be seen in the first part of the table (steady GP) where ICC is noticeably larger. The second point is concerned with ICC estimates themselves. In both cases, the GEE1.5 estimate of ρ_{HH} is substantially smaller than its GLMM counterpart. This observation clearly necessitates further research and it would be interesting to ascertain whether a link can be established between these two types of estimates.

TABLE 3
Parameter estimates for the binary response models
(standard errors are given in parentheses)

	Steady GP				Subjective health			
Parameter	GEE1.5		GLMM		GEE1.5		GLMM	
β Intercept	1.276	(0.292)	3.506	(0.718)	1.135	(0.203)	1.347	(0.254)
Sex (Male)	− 0.309	(0.051)	− 0.727	(0.124)	0.339	(0.056)	0.409	(0.068)
Age†								
15-24	0.220	(0.104)	0.534	(0.252)	–		–	
25-34	0.162	(0.083)	0.394	(0.183)	− 0.245	(0.139)	− 0.271	(0.158)
35-44	0.236	(0.083)	0.576	(0.184)	− 0.981	(0.125)	− 1.138	(0.142)
45-54	0.447	(0.108)	0.058	(0.255)	− 1.284	(0.126)	− 1.485	(0.144)
55-64	0.680	(0.137)	1.565	(0.306)	− 1.617	(0.130)	− 1.891	(0.154)
65-74	1.022	(0.159)	2.246	(0.338)	− 1.865	(0.131)	− 2.193	(0.157)
75+	1.319	(0.204)	2.966	(0.434)	− 2.152	(0.143)	− 2.544	(0.173)
Income (reference: < 20,000)								
20,000-30,000	− 0.081	(0.172)	− 0.142	(0.399)	− 0.167	(0.128)	− 0.220	(0.155)
30,000-40,000	0.544	(0.183)	1.261	(0.416)	0.076	(0.127)	0.079	(0.153)
40,000-60,000	0.509	(0.173)	1.210	(0.396)	0.335	(0.127)	0.398	(0.154)
> 60,000	0.328	(0.192)	0.691	(0.446)	0.761	(0.159)	0.895	(0.192)
Education (reference: No diploma)								
Primary	− 0.295	(0.295)	− 0.985	(0.715)	0.158	(0.164)	0.182	(0.206)
Lower secondary	− 0.505	(0.285)	− 1.465	(0.698)	0.406	(0.163)	0.494	(0.206)
Higher secondary	− 0.253	(0.279)	− 0.808	(0.689)	0.743	(0.161)	0.903	(0.205)
Higher	− 0.805	(0.276)	− 2.106	(0.689)	1.060	(0.166)	1.254	(0.210)
Region (reference: Brussels)								
Flanders	1.561	(0.126)	3.598	(0.316)	0.434	(0.082)	0.507	(0.097)
Wallonia	1.338	(0.113)	3.060	(0.283)	− 0.094	(0.076)	− 0.115	(0.092)
σ^2_u	–		15.09	(1.779)	–		1.160	(0.202)
ρ_{HH}††	0.596	(0.027)	0.821	(0.062)	0.136	(0.018)	0.261	(0.022)

† Reference category is 0-14 for steady GP and 14-25 for subjective health.
†† Intra-class correlation.

5. Discussion

In this paper we have considered multilevel modeling techniques to investigate the effect of clustering in the HIS. This was motivated by the fact that these methods fully acknowledge and take advantage of the underlying hierarchical structure present in the data, thereby dealing with the clustering aspects of the HIS in an efficient manner. In addition, from a model-based perspective we are able to incorporate design-related information directly into the model to adjust for the effects of the sampling design.

While sampling schemes are commonly ignored in multilevel analyses of survey data since it is conceivable, at least theoretically, to include the whole design-related information in the model, this may not always be feasible. It is therefore useful to examine weighting in multilevel models to adjust for the effects of sampling in such cases. In this paper we have utilized a simple and easy-to-implement procedure which was recently proposed by Pfeffermann *et al.* (1) as an alternative to their fully-integrated weighting procedure, although they recommend to use it with caution as it can give biased results in some circumstances.

Due to software limitations, we failed to calculate accurate standard errors for the weighted estimator and this was reflected in greatly inflated estimates compared to the unweighted scenario. Whether or not this is due to the empirical estimator itself, this calls for better, appropriate software which would enable one to perform this kind of analyses routinely. It should be noticed that the full weighting procedure proposed in the aforementioned paper will be implemented in the next MLwiN release, thereby eliminating the above concerns and alleviating considerably the computational burden.

Assuming we have a fully-integrated weighting procedure at our disposal, we can wonder whether "to weight or not to weight". On the one hand, if one is willing to include the whole information related to the sampling design, the resulting analysis will be safe. On the other hand, it will doubtlessly be difficult, if not impossible, to do so in practice and one should then consider weighting as an alternative. Note that screening design information for possible inclusion in the model is certainly a good statistical practice and should be examined whenever it is feasible. Moreover, a comparison between weighted and unweighted estimators would be most relevant, with large discrepancies calling for caution.

Another point related to weighting concerns model misspecification. Indisputably, we can be quite confident that the models we contemplated in Section 4 were actually misspecified. It has been argued that weighting can protect against model misspecification (30), and one can therefore wonder whether such a property carries over to multilevel modeling. Pfeffermann *et al.* (1), however, assume their model to be correctly specified and do not explore the issue of model misspecification in their paper. This topic would benefit from further research. It should also be emphasized that weights have not been adjusted for non-response nor poststratified in the present paper. The effect of non-response in the HIS is addressed in a second paper (31). It is planned to investigate the combined influence of clustering and non-response in the near future.

From a practitioner's point of view one may ask how does a multilevel analysis, whether weighted or unweighted, compare with the more traditional, design-based approach. The following may serve as a rough guideline. While an unweighted multilevel analysis may seem slightly more efficient and produces consistent parameter estimates, there might be problems with the adequacy of the precision measures. Both a design-based analysis (as typically carried out in STATA or SUDAAN) as well as a weighted multilevel analysis are adequate. The former is fine if we are merely interested in obtaining correct inference about main effects (such as age, sex, or country effects), whereas the second one is necessary if there is at least some interest in the clustering effects themselves. Finally, if a multilevel analysis is unavailable, e.g., due to software unavailability, then a classical design-based analysis may be the only option.

Finally, the above discussion deals with the multilevel linear model assuming response data to be normally distributed. It is by far less straightforward to treat the case of generalized linear models (GLIM) in general, and binary response models in particular, from a multilevel perspective. Besides genuine estimation problems typically encountered when random effects are being introduced in a GLIM framework, there still lacks a thorough investigation of weighting in this setting. This issue definitely merits further exploration. Furthermore, a better understanding of the marginal and random-effects approaches to estimating intra-class correlation would be useful.

References

1. PFEFFERMANN D, SKINNER C J, HOLMES D J, GOLDSTEIN H AND RASBASH J. Weighting for unequal selection probabilities in multilevel models. J R Stat Soc B 1998, 60: 23-40.
2. GRAUBARD B I AND KORN E L. Modelling the sampling design in the analysis of health surveys. Stat Meth Med Res 1996, 5: 263-281.
3. SKINNER C J, HOLT D AND SMITH D M F. Analysis of Complex Surveys. Chichester: Wiley, 1989.
4. GOLDSTEIN H. Multilevel Statistical Models, 2nd edition. London: Edward Arnold, 1995.
5. QUATAERT P, VAN OYEN H, TAFFOREAU J, SCHIETTECATTE L, LEBRUN L, BELLAMAMMER L, MOLENBERGHS G. Health Interview Survey, 1997: Protocol for the selection of the households and the respondents. Brussels: SPH/EPISERIE N12, 1997.
6. VAN OYEN H, TAFFOREAU J, HERMANS H, QUATAERT P, SCHIETTECATTE E, LEBRUN L, BELLAMAMMER L. The Belgian Health Interview Survey. Arch Public Health 1997, 52: 79-82.
7. BRYK A S AND RAUDENBUSH S W. Hierarchical Linear Models: Applications and Data Analysis Methods. Newbury Park: Sage Publications, 1992.

8. LONGFORD N T. Random Coefficient Models. Oxford University Press: Oxford, 1993.
9. KREFT I AND DE LEEUW J. Introducing Multilevel Modeling. London: Sage Publications, 1998.
10. SEARLE S R, CASELLA G AND MCCULLOGH C E. Variance Components. New York: John Wiley & Sons, 1992.
11. VERBEKE G AND MOLENBERGHS G. Linear Mixed Models in Practice: A SAS Oriented Approach. New York: Springer-Verlag, 1997.
12. GOLDSTEIN H. Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika 1986, 73: 43-56.
13. LONGFORD N T. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. Biometrika 1987, 74: 812-827.
14. RAUDENBUSH S W AND BRYK A S. A hierarchical model for studying school effects. Sociol Educ 1986, 59: 1-17.
15. LIANG K Y AND ZEGER S L. Longitudinal data analysis using generalised linear models. Biometrika 1986, 73: 13-22.
16. Mc CULLAGH P AND NELDER J. Generalised Linear Models, 2nd edition. London: Chapman and Hall, 1989.
17. ZEGER S L AND KARIM R M. Generalized linear models with random effects: a Gibbs sampler approach. J Am Stat Assoc 1991, 86: 79-86.
18. RODRÍGUEZ G AND GOLDMAN N. An assessment of estimation procedures for multilevel models with binary responses. J R Stat Soc A 1995, 158: 73-89.
19. GOLDSTEIN H. Nonlinear multilevel models with an application to discrete response data. Biometrika 1991, 78: 45-51.
20. BRESLOW N E AND CLAYTON D G. Approximate inference in generalized linear model. J Am Stat Assoc 1993, 88: 9-25.
21. GOLDSTEIN H AND RASBASH J. Improved approximations for multilevel models with binary responses. J R Stat Soc A 1996, 159: 505-513.
22. WOLFINGER R AND O'CONNELL M. Generalized linear mixed models: A pseudo-likelihood approach. J Statist Comput Simul 1993, 48: 233-243.
23. GOLDSTEIN H, RASBASH J, PLEWIS I, DRAPER D, BROWNE W, YANG M, WOODHOUSE G AND HEALY M. A user's guide to MLwiN. London: Institute of Education, 1998.
24. HEDEKER D AND GIBBONS R D. MIXOR. A Computer Program for Mixed-Effects Ordinal Probit and Logistic Regression Analysis. University of Illinois at Chicago, 1993.
25. KORN E L AND GRAUBARD B I. Analysis of large health surveys. J R Stat Soc A 1995, 158: 263-295.
26. HERSON J. Fieller's theorem vs. the delta method for significance intervals for ratios. J Statist Comput Simul 1975, 3: 265-274.
27. HEDEKER D AND GIBBONS R D. A random-effects ordinal regression model for multilevel analysis. Biometrics 1994, 50: 933-944.
28. Prentice R L Correlated binary regression with covariates specific to each binary observation. Biometrics 1988, 44: 1033-1048.
29. NEUHAUS J M, KALBFLEISCH J D AND HAUCK W W. A comparison of cluster-specific and population averaged approaches for analyzing correlated binary data. Int Statist Rev 1991, 59: 25-36.
30. PFEFFERMANN D. The role of sampling weights when modelling survey data. Int Statist Rev 1993, 61: 317-337.
31. BURZYKOWSKI T, MOLENBERGHS G, TAFFOREAU J, VAN OYEN H, DEMAREST S AND BELLAMMAMER L. Missing data in the Health Interview Survey in Belgium. Submitted, 1998.