STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:4045–4054
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2386

Accommodation of additional non-randomly sampled cases in a study of *Helicobacter pylori* infection in families

Mårten Kivi[1,2,‡], Anna L. V. Johansson[1,*,†,‡], Agus Salim[3], Ylva Tindberg[1] and Marie Reilly[1]

[1]*Department of Medical Epidemiology and Biostatistics* (*MEB*), *Karolinska Institutet*, *Stockholm*, *Sweden*
[2]*Department of Clinical Microbiology*, *Microbiology and Tumor Biology Center* (*MTC*), *Karolinska Institutet*, *Stockholm*, *Sweden*
[3]*National Centre for Epidemiology and Population Health*, *The Australian National University*, *Canberra*, *Australia*

SUMMARY

Epidemiological studies with two-stage designs typically gather information about some covariates from all study subjects in the first sampling stage, while additional data from only a subset of the subjects are collected in the second sampling stage. Appropriate analysis of two-stage studies maintains validity and can also improve precision. We describe an application of a weighted likelihood method, mean-score logistic regression, to accommodate data from a cross-sectional study of *Helicobacter pylori* infection in children, where the study sample was enriched with additional non-randomly sampled cases. The present work exemplifies how careful analysis of epidemiological data from complex sampling schemes can adjust for potential selection bias, improve precision and enable a more complete investigation of factors of interest. Our results highlight the importance of *H. pylori* infected mothers and siblings as risk factors for the infection in children in Sweden. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: mean-score logistic regression; two-stage; missing data; *Helicobacter pylori*

1. INTRODUCTION

Random samples are fundamental to inference in epidemiological studies and unintentional non-random missingness may introduce unforeseen bias. Such bias can however be avoided

*Correspondence to: Anna L. V. Johansson, MEB, Box 281, Karolinska Institutet, 171 77 Stockholm, Sweden.
†E-mail: anna.johansson@ki.se
‡These authors contributed equally to this work.

Contract/grant sponsor: The Swedish Cancer Society; contract/grant number: 4858-B03-01XAB
Contract/grant sponsor: Karolinska Institutet
Contract/grant sponsor: Sven Jerring Foundation
Contract/grant sponsor: Swedish Medical Society
Contract/grant sponsor: Goljes Foundation
Contract/grant sponsor: Foundation Samariten

Received 18 August 2005
Accepted 31 August 2005

when the determinants for the missingness are known, thus allowing the non-randomness to be accounted for [1–4]. Furthermore, systematically skewed sampling may be preferred over random sampling to improve statistical efficiency when economic or practical restrictions are placed on the amount of data that can be collected [5, 6]. Intentionally skewed sampling schemes should however be carefully planned and an appropriate analysis must be applied to ensure the validity of the parameter estimates.

Studies with missing data where the determinants of the missingness are identifiable can be considered as two-stage sampling schemes, regardless of whether the two-stage nature of the data has arisen unintentionally or by design. Typically, the first sampling stage gathers information about some covariates from all study subjects, while the second sampling stage collects additional data from only a subset of the subjects. Statistical methods for analysis of two-stage studies include imputation [7], non-parametric weighted likelihood [8] and pseudo-likelihood and full maximum likelihood approaches [9].

We describe herein an application of a weighted likelihood method to accommodate data from a cross-sectional study of *Helicobacter pylori* infection in children, where the study sample was enriched with additional non-randomly sampled cases. The gastric bacterium *H. pylori* is one of the most common human infections worldwide and the prevalence in child populations ranges from under 10 per cent in high-income countries to over 80 per cent in low-income countries [10]. *H. pylori* infection may persist for decades and contributes substantially to the development of peptic ulcer disease and gastric cancer [11]. The transmission of *H. pylori* is inadequately understood but the infection is typically acquired in early childhood [12]. There are no consistent and verified environmental reservoirs, and intrafamilial person-to-person transmission appears to predominate [13, 14]. A pattern of transmission from mothers to children, between siblings but not from fathers to children has been discerned [13, 15–19]. This requirement for prolonged intimate contact is in line with a relatively low infectiousness of the bacterium.

A cross-sectional serological survey in 10–12-year-old school children in Stockholm identified the family as a major contributor to *H. pylori* transmission [13]. A primary risk factor for infection in the children was family origin in countries with high *H. pylori* prevalence. Household socioeconomic factors also played a role while *H. pylori* infection in classmates was not found to be a risk factor. Prompted by these findings, a second sampling stage was planned, where *H. pylori* infection status of the different family members was assessed to help clarify their contribution to infection in a subset of the children. Having an infected mother, infected siblings and being born in a high-prevalence country were all identified as strong markers of risk for infection in these children [19]. However, the children sampled in the second-stage deliberately included some additional non-randomly sampled cases who were omitted from that analysis. The objective of the present work was to investigate if appropriate analysis of all available data, including the additional cases, would reveal further insights into the familial clustering of *H. pylori* infection, while possibly improving the precision of the estimates of risk factors already identified.

2. MATERIAL AND METHODS

2.1. Data collection

The first-stage of this study sampled 679 10–12-year-old children in a cross-sectional *H. pylori* serological survey in 11 Stockholm schools (schools A–K) conducted between February and

Table I. Participation of *H. pylori* infected index children (cases) and uninfected index children in schools A–D and E–K.

	Infected index children		Uninfected index children	
School	First-stage*, *n*	Second-stage†, *n* (per cent)	First-stage*, *n*	Second-stage†, *n* (per cent)
A–D	68	54 (79)	165	108 (65)
E–K	37	33 (89)	409	Not sampled
Total	105	87 (83)	574	108 (19)

*Children in the initial serological school survey.
†Index children whose family members contributed a blood sample and answered a questionnaire in the second sampling stage.

April 1998 [13]. In the second sampling stage between November 1998 and May 1999, family members residing in the household at least 14 days per month were invited to complete questionnaires and contribute blood for serological infection status determination [19]. The school children served as infected index children (cases) and uninfected index children while the family members represented the exposure of interest. For families where two children were sampled at the first-stage ($n = 10$), one child was included as an index child and the other contributed exposure as a sibling.

The sampling of index children into the second stage was not random. Households of low socioeconomic status (SES) and immigrant background were preferentially targeted because these factors are associated with *H. pylori* infection and hence these families were regarded as most informative. All index children from the four schools (schools A–D) with the highest *H. pylori* prevalence were invited to participate while only infected index children were invited from the remaining seven schools (schools E–K). Index children were included in the second-stage sample when at least one family member contributed questionnaire information and a blood sample. The participation rate of invited children in the second-stage was 87 of 105 (83 per cent) for infected index children and 108 of 165 (65 per cent) for uninfected index children (Table I). Of the 664 identified family members, 582 (88 per cent) contributed a blood sample and questionnaire information. The local ethics committee approved the study protocol and children and/or parents gave informed consent.

Family members were classified as uninfected, infected or absent from the household. Household SES was categorized as low (blue-collar worker, self-employed, unemployed) or high (white-collar worker) [20] and antibiotic consumption as 0–5 or $\geqslant 6$ courses in life. The countries of birth of index children and parents were classified as low-prevalence areas (Western Europe and North America) or high-prevalence areas (Middle East, Turkey, Eastern Europe, Africa, Asia, Latin America and South America) [10, 13].

2.2. Statistical methods

The associations between exposures and *H. pylori* infection in the index children were estimated with odds ratios (OR) and 95 per cent confidence intervals (CI). Naïve logistic regression and mean-score logistic regression (http://www.meb.ki.se/~marrei/software/, accessed 17 June 2005) were performed in the STATA (version 8.0, Stata Corporation, TX, USA, http://www.stata.com) and R packages (version 1.9.0, R Foundation for Statistical

Computing, http://www.R-project.org). For comparison of efficiency, the pseudo-likelihood approach by Breslow and Holubkov [9] was implemented using a modified version of their S-Plus software (http://faculty.washington.edu/norm/software.html, accessed 17 June 2005) run in the R package.

Reilly and Pepe introduced mean-score logistic regression as a method to obtain valid estimates of ORs and standard errors in two-stage studies with complete first-stage and incomplete second-stage covariate information. Unbiased effect estimates of the second-stage covariates are obtained by a weighted likelihood method, where the weights are calculated from the sampling fractions in strata specified by first-stage covariates that are determinants of participation in the second-stage [8]. Denoting the outcome variable as Y, the complete covariates in the first-stage as Z and the incomplete covariates in the second-stage as X, then the mean-score method solves the score equation:

$$\sum_i \frac{n^{Z_i Y_i}}{n_2^{Z_i Y_i}} S_\beta(Y_i|X_i, Z_i) = 0$$

where $n^{Z_i Y_i}$ and $n_2^{Z_i Y_i}$ represent the total number of observations and the number of second-stage (i.e. complete) observations with $Z = Z_i$ and $Y = Y_i$, and summation is over all second-stage observations. Thus, mean-score performs a weighted likelihood analysis with the weights $n^{Z_i Y_i}/n_2^{Z_i Y_i}$ in the strata defined by Y and Z. For the logistic regression model, the score $S_\beta(Y_i|X_i, Z_i)$ reduces to

$$S_\beta(Y_i|X_i, Z_i) = [y_i - \theta_i] \cdot \begin{bmatrix} 1 \\ x_i \\ z_i \end{bmatrix}$$

where

$$\theta_i = \text{prob}(y_i = 1) = \frac{e^{\beta_0 + \beta_1 x_i + \beta_2 z_i}}{1 + e^{\beta_0 + \beta_1 x_i + \beta_2 z_i}}$$

The asymptotic variance of the mean-score estimator is estimated by

$$\frac{I^{-1} + I^{-1}\Omega I^{-1}}{n}$$

where n is the total number of observations, I is the usual information matrix for second-stage (i.e. complete) observations and Ω is a weighted sum of the variances of the scores in the various Y, Z strata, estimated by

$$\hat{\Omega} = \sum_i \frac{n^{Z_i Y_i}}{n} \frac{n_1^{Z_i Y_i}}{n_2^{Z_i Y_i}} \hat{\text{var}}[S_\beta(Y_i|X_i, Z_i)|Y_i, Z_i]$$

where $n_1^{Z_i Y_i}$ is the number of first-stage observations with $Z = Z_i$ and $Y = Y_i$, i.e. $n^{Z_i Y_i} = n_1^{Z_i Y_i} + n_2^{Z_i Y_i}$. The summation for the above formula is taken over all unique combinations of Z and Y.

Due to the conditioning on Y and Z, mean-score accommodates data missing-at-random (MAR) within strata defined by the first-stage covariates and the outcome. This implies that within these strata available observations at the second-stage are representative of the relationships between outcome, predictors and confounders.

3. APPLICATION AND RESULTS

3.1. First-stage variables in the mean-score analysis

Sociodemographic characteristics of the 679 infected and uninfected children in the cross-sectional serological school survey are depicted in Figure 1. Low household SES and origin in countries with high *H. pylori* prevalence were more common for infected children and for children from schools A–D, as compared to uninfected children and children from schools E–K. The skewed sampling scheme rendered the second-stage sample unrepresentative of all first-stage observations (Table I). School could not be used as a first-stage covariate in the mean-score analyses to control for bias due to the sampling scheme because no uninfected index children were sampled from schools E–K. Instead, we explored using as first-stage covariates the sociodemographic variables that prompted the biased sampling scheme.

Factors associated with participation in the second-stage were assessed for index children by logistic regression models, with the outcome being second-stage participation. The high response rate of infected index children from all schools resulted in this population being well represented in our sample. In logistic regression models, none of the investigated factors (SES, gender, family size, country of birth of index children and parents) were associated with second-stage participation by infected index children. Second-stage participation among uninfected index children was found to be independently associated with having at least one parent born in a high-prevalence country (OR 2.1, 95 per cent CI 1.2–3.5) and low SES (OR 1.9, 95 per cent CI 1.2–2.9), adjusted for the child's country of birth (OR 1.1, 95 per cent CI 0.5–2.3). Gender and family size were unrelated to inclusion in the second-stage sample for uninfected index children. Thus, the second-stage data for uninfected index children were regarded as a random sample within the strata defined by SES and the country of birth of the parents.

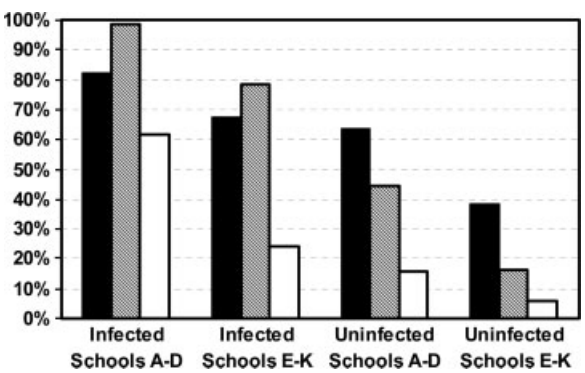

Figure 1. Sociodemographic characteristics of the 679 children in the serological school survey, i.e. the first-stage sample, are presented stratified by the *H. pylori* infection status of the children and the schools that were differentially sampled in the second stage. All children in schools A–D were invited to participate in the second-stage sample while only infected children were invited from schools E–K. The bars show the percentages of children who come from households of low SES (black), have at least one parent born in a high-prevalence country (striped) and are born in a high-prevalence country themselves (white).

Accordingly, the dichotomous variables for household SES and parents' place of birth were specified as 'first-stage' variables in the call to the mean-score function. This decision was based on the *a priori* knowledge that households of low SES and immigrant background were targeted for second-stage sampling, and was supported by the analyses above, which identified no additional predictors for second-stage participation. Thus, in our weighted likelihood Y is the infection status of the index child, X is the infection status of the family members and Z is the household SES and parents' place of birth. The MAR assumption translates into index children with second-stage data being assumed to be similar to index children without second-stage data, within the strata defined by the outcome, household SES and parents' place of birth.

3.2. Comparison of naïve logistic regression and mean-score logistic regression

In Table II, we present the results of mean-score logistic regression in all schools (A–K) and compare to a naïve analysis of all available data from these schools. For comparison, a naïve logistic regression analysis was also performed for schools A–D, where all families of infected and uninfected index children had been invited to participate in the second sampling stage. Comparing the two naïve analyses, there was an expected gain in precision due to the larger sample sizes when using data from all schools, but changes in the point estimates reflect potential bias from the sampling scheme.

The mean-score logistic regression analysis includes observations from all schools without impairing validity and enables some additional risk factors to be examined. For example, the effect of an absent mother could not be estimated in the analysis restricted to schools A–D. The mean-score analysis of all schools suggested that children of an absent mother were at higher risk of infection compared to children living with an uninfected mother (OR 3.8, 95 per cent CI 0.4–32.9) (Table II). In the analysis of schools A–D, a similar trend was observed for children without siblings compared to children with only uninfected siblings (OR 5.2, 95 per cent CI 0.7–38.1). This finding was corroborated in the mean-score analyses of all schools, where the gain in precision from the larger sample size resulted in a statistically significant effect for absent siblings (OR 6.1, 95 per cent CI 1.3–27.7). The contribution of infected siblings to infection in the index children was further assessed for younger and older siblings, respectively, and no difference could be detected although the point estimates differed slightly (mean-score adjusted OR 16.0, 95 per cent CI 3.3–78.6 and 12.2, 95 per cent CI 1.8–81.0 for younger and older siblings, respectively).

Extending the analysis to all schools also made it possible to perform an analysis of only biological nuclear families by excluding index children with at least one step-parent ($n = 22$). In this analysis, the importance of having an infected mother compared to an uninfected mother tended to increase although the precision was low (mean-score adjusted OR 70.9, 95 per cent CI 7.2–701.4). Furthermore, the estimate for having an absent mother increased and reached statistical significance (mean-score adjusted OR 18.5, 95 per cent CI 1.1–308.1).

To assess the presence of confounding by school or class, results from unconditional logistic regression were compared with results from conditional logistic regression with school or class as the matching factors. There was no evidence of confounding by school or class with similar effect estimates being obtained from the unconditional and conditional logistic regression. Including prevalence of *H. pylori* in classmates in our unconditional model, we confirmed the previous finding that the *H. pylori* prevalence in classmates was not a risk

Table II. *H. pylori* infection status in different categories of family members as risk factors for the infection in the index children.

	Schools A–D			Schools A–K				
	H. pylori infected		Naïve logistic regression, adj. OR	*H. pylori* infected		Naïve logistic regression, adj. OR	Pseudo-likelihood logistic regression, adj. OR	Mean-score logistic regression, adj. OR
Variable	*n*/total	per cent	(95 per cent CI)*	*n*/total	per cent	(95 per cent CI)*	(95 per cent CI)*	(95 per cent CI)*
Mother								
Uninfected	3/66	5	1.0	8/71	11	1.0	1.0	1.0
Infected	47/82	57	11.6 (2.0–67.9)	71/106	67	9.6 (2.7–34.5)	14.0 (3.8–51.7)	12.8 (3.3–49.1)
Absent mother	0/8	0	NA†	2/10	20	4.2 (0.5–35.8)	4.3 (0.6–32.9)	3.8 (0.4–32.9)
Father								
Uninfected	3/44	7	1.0	12/53	23	1.0	1.0	1.0
Infected	32/61	52	1.4 (0.2–9.8)	47/76	62	1.4 (0.4–5.1)	1.9 (0.5–6.9)	1.8 (0.5–6.6)
Absent father	14/40	35	1.4 (0.2–10.8)	21/47	45	1.0 (0.2–4.4)	1.3 (0.3–5.5)	1.4 (0.3–6.7)
Siblings								
All siblings uninfected‡	10/73	14	1.0	15/78	19	1.0	1.0	1.0
$\geqslant 1$ infected sibling	34/44	77	8.1 (1.8–37.3)	54/64	84	11.1 (3.3–37.5)	11.1 (3.0–41.1)	10.3 (2.8–38.3)
Absence of siblings	5/20	25	5.2 (0.7–38.1)	9/24	38	6.4 (1.5–28.2)	7.0 (1.7–29.4)	6.1 (1.3–27.7)

CI: confidence interval.
OR: odds ratio.
*Adjusted for all variables in the table, household socioeconomic status, the index child's place of birth and antibiotic consumption.
†Not applicable. There were no infected children without a mother in the household.
‡All children in the sibships that fulfilled the inclusion criterion participated in the study and were uninfected.

factor for infection in the index children [13]. Furthermore, household SES and country of birth of classmates were not related to index child infection, neither before nor after restriction to classes with a *H. pylori* prevalence of at least 10 per cent.

4. DISCUSSION

The present application of mean-score logistic regression to accommodate a non-random two-stage sampling scheme enabled a more complete analysis of familial risk factors for *H. pylori* infection in children. In a naïve analysis of all second-stage data, the uninfected children would be expected to have an overrepresentation of infected family members due to the sampling scheme targeting families of low SES and immigrant background. This could lead to an underestimation of the risk associated with having infected family members. A tendency in this direction was observed for the contribution of an infected mother to the index child's risk in the naïve analysis of schools A–K. This indicates that the mean-score method offers a more appropriate analysis of all the available data.

The improved precision in the analyses of all schools compared to schools A–D could largely be attributed to the inclusion of the 33 additional cases rather than to the mean-score method *per se*. We had anticipated that mean-score would improve precision compared to naïve logistic regression of only second-stage data, because mean-score also considers the information in the first-stage sample. This precision benefit should be most pronounced when there is a substantial proportion of missing data. However, when data become too sparse, the standard errors are compromised by the smallest strata and there may be little or no overall gain in precision as occurred here. Since the infected index children could reasonably be regarded as a random sample within all schools they did not necessarily require reweighting for validity purposes. To reduce the level of stratification in the data, we repeated our analysis using a five-level first-stage variable taking one value for infected index children and four values based on household SES and the parents' country of birth for uninfected index children. The results of this alternative analysis were similar to those presented with no appreciable gain in precision. Furthermore, a pseudo-likelihood analysis yielded similar estimates with only marginal improvements in precision (Table II). Thus, the mean-score is a reasonable choice of method here based on its performance and simplicity. Further gains in precision could have been realized if the additional second-stage observations had been sampled according to some more optimal design [5]. We found that such a design should have included more uninfected index children from households of low SES and with parents born in high-prevalence regions.

When selecting first-stage variables for mean-score analysis, including factors that predict participation in the second-stage should maintain validity. There will be greater precision in the second-stage effect estimates if the chosen first-stage variables are correlated with the second-stage variables of interest, as then the first-stage incomplete subjects contribute more 'information' [3]. The present sampling scheme targeted households of low SES and immigrant background and these variables were used as first-stage variables to control for bias due to the sampling. The optimal predictor for participation in the second-stage of uninfected index children would be whether they attended schools A–D or schools E–K. However, school could not be used as a first-stage variable because no uninfected index children were sampled from schools E–K due to the manner in which schools served as the sampling frame to obtain families with the desired characteristics. Means to accommodate zero-sampling-fractions have

been proposed but rely on additional assumptions [21] and are not required here since we can use the strata defined by SES and immigrant background as first-stage variables, provided that there is no confounding by school or class. There was no evidence of confounding by school or class when comparing results from unconditional and conditional logistic regression with school or class as the matching factors. Furthermore, the *H. pylori* prevalence and sociodemographic characteristics of classmates were refuted as risk factors for the infection in the present population of children, consistent with several studies that point to the family as the predominant framework for transmission. In light of these findings, mechanisms that could introduce substantial confounding by school or class after considering household SES and immigrant background are improbable in the present study. Nevertheless, interpretation of results obtained by missing data applications should always take the strong (and by definition not testable) MAR assumption into consideration.

The major benefit of the re-analysis of this data is that exposures that had not been previously studied could be investigated due to the inclusion of the additional cases. The results of a standard logistic regression analysis of the second-stage data in schools A–D are presented and discussed in more detail elsewhere [19]. Briefly, having an *H. pylori* infected mother or infected siblings and being born in a high-prevalence country were primary independent risk factors for infection in the children, which illustrates the importance of both current familial infections and country of origin. The present analyses include novel modelling approaches that consider the infection status of different categories of family members and despite the low precision of some estimates the analyses allow us to extend the interpretation of previous findings.

The associations between infection in the index children and absence of a mother or siblings in the household may be explained by previous or intermittent presence of these potentially infectious family members, who do not fulfill the inclusion criterion for this study. Furthermore, the association between infection in mothers and index children tended to increase when only biological mothers were considered. This finding might be attributable to a biological mother being a source of infection since the birth of the index child, while a stepmother is more likely to have entered the family at a later stage. We found no difference between the contributions of younger and older infected siblings. A previous study suggested older siblings to be more important, but these findings may not be comparable because that work was performed in a high-prevalence setting and did not adjust for parental infection status [16].

In conclusion, the present work exemplifies how careful analysis of epidemiological data from complex sampling schemes can adjust for potential selection bias, improve precision and enable a more complete investigation of factors of interest. Our results highlight the importance of *H. pylori* infected mothers and siblings as risk factors for the infection in children in Sweden.

ACKNOWLEDGEMENTS

We thank Professor Marta Granström and Carina Bengtsson for the serological analyses that defined the infection status of the participants.

REFERENCES

1. Flanders WD, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine* 1991; **10**(5):739–747.

2. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. *Biometrika* 1988; **75**(1):11–20.
3. Reilly M, Sullivan Pepe M. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* 1995; **82**(2):299–314.
4. Zhao LP, Lipsitz S. Designs and analysis of two-stage studies. *Statistics in Medicine* 1992; **11**(6):769–782.
5. Reilly M. Optimal sampling strategies for two-stage studies. *American Journal of Epidemiology* 1996; **143**(1): 92–100.
6. Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. *American Journal of Epidemiology* 1988; **128**(6):1198–1206.
7. Little RJ, Rubin DB. *Statistical Analysis with Missing Data* (2nd edn). Wiley: New Jersey, U.S.A., 2002.
8. Reilly M, Pepe M. The relationship between hot-deck multiple imputation and weighted likelihood. *Statistics in Medicine* 1997; **16**(1–3):5–19. DOI: 10.1002/(SICI)1097-0258(19970115)16:1<5::AID-SIM469>3.0.CO;2-8
9. Breslow NE, Holubkov R. Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine* 1997; **16**(1–3):103–116. DOI: 10.1002/(SICI)1097-0258(19970115)16:1<103::AID-SIM474>3.0.CO;2-P
10. Torres J, Perez-Perez G, Goodman KJ, Atherton JC, Gold BD, Harris PR, la Garza AM, Guarner J, Munoz O. A comprehensive review of the natural history of *Helicobacter pylori* infection in children. *Archives of Medical Research* 2000; **31**(5):431–469. DOI: 10.1016/S0188-4409(00)00099-0
11. Suerbaum S, Michetti P. *Helicobacter pylori* infection. *The New England Journal of Medicine* 2002; **347**(15): 1175–1186.
12. Granström M, Tindberg Y, Blennow M. Seroepidemiology of *Helicobacter pylori* infection in a cohort of children monitored from 6 months to 11 years of age. *Journal of Clinical Microbiology* 1997; **35**(2):468–470.
13. Tindberg Y, Bengtsson C, Granath F, Blennow M, Nyrén O, Granström M. *Helicobacter pylori* infection in Swedish school children: lack of evidence of child-to-child transmission outside the family. *Gastroenterology* 2001; **121**(2):310–316. DOI: 10.1053/gast.2001.26282
14. Mendall MA, Goggin PM, Molineaux N, Levy J, Toosy T, Strachan D, Northfield TC. Childhood living conditions and *Helicobacter pylori* seropositivity in adult life. *The Lancet* 1992; **339**(8798):896–897.
15. Rocha GA, Rocha AM, Silva LD, Santos A, Bocewicz AC, Queiroz Rd Rde M, Bethony J, Gazzinelli A, Correa-Oliveira R, Queiroz DM. Transmission of *Helicobacter pylori* infection in families of preschool-aged children from Minas Gerais, Brazil. *Tropical Medicine and International Health* 2003; **8**(11):987–991. DOI: 10.1046/j.1360-2276.2003.01121.x
16. Goodman KJ, Correa P. Transmission of *Helicobacter pylori* among siblings. *The Lancet* 2000; **355**(9201): 358–362. DOI: 10.1016/S0140-6736(99)05273-3
17. Rothenbacher D, Winkler M, Gonser T, Adler G, Brenner H. Role of infected parents in transmission of *Helicobacter pylori* to their children. *The Pediatric Infectious Disease Journal* 2002; **21**(7):674–679. DOI: 10.1097/01.inf.0000021081.69738.42
18. Kivi M, Tindberg Y, Sörberg M, Casswall TH, Befrits R, Hellström PM, Bengtsson C, Engstrand L, Granström M. Concordance of *Helicobacter pylori* strains within families. *Journal of Clinical Microbiology* 2003; **41**(12):5604–5608. DOI: 10.1128/JCM.40.6.2192-2198.2002
19. Kivi M, Johansson ALV, Reilly M, Tindberg Y. *Helicobacter pylori* status in family members as risk factors for infection in children. *Epidemiology and Infection*, in press. DOI: 10.1017/S0950268805003900
20. Statistics Sweden. *Swedish Socioeconomic Classification. Reports on Statistical Co-ordination 1982:4* (Swedish). Statistics Sweden: Örebro, Sweden, 1995.
21. Chatterjee N, Chen Y-H, Breslow NE. A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association* 2003; **98**(461):158–168. DOI: 10.1198/016214503388619184

