Combining data from 2 nested case–control studies of overlapping cohorts to improve efficiency

AGUS SALIM∗

Department of Community, Occupational and Family Medicine, Yong Loo Lin School of Medicine, National University of Singapore and Centre for Mental Health Research, Australian National University Canberra ACT 0200, Australia

CHRISTINA HULTMAN, PÅR SPARÉN, MARIE REILLY

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet Box 281 SE 17177, Stockholm, Sweden
cofaguss@nus.edu.sg

SUMMARY

Researchers subject to time and budget constraints may conduct small nested case–control studies with individually matched controls to help optimize statistical power. In this paper, we show how precision can be improved considerably by combining data from a small nested case–control study with data from a larger nested case–control study of a different outcome in the same or overlapping cohort. Our approach is based on the inverse probability weighting concept, in which the log-likelihood contribution of each individual observation is weighted by the inverse of its probability of inclusion in either study. We illustrate our approach using simulated data and an application where we combine data sets from 2 nested case–control studies to investigate risk factors for anorexia nervosa in a cohort of young women in Sweden.

Keywords: Anorexia; Cost efficiency; Matching; Proportional hazards; Weighted likelihood.

1. INTRODUCTION

The case–control design is one of the most widely used study designs in epidemiology and is especially attractive for low prevalence diseases where a large cohort would need to be followed over a long time period to obtain precise estimates of the effects of risk factors. The sampling procedure in a case–control design is intentionally biased. The probability of being selected into the study is not the same for all individuals but rather depends on their status with respect to the disease studied and, in matched case–control studies, on their matching variables. Despite this biased sampling, it is well known that ordinary logistic regression (or in the case of matched case–control designs, conditional logistic regression) can be used to obtain unbiased estimates of the odds ratios for risk factors of interest (Breslow and Day, 1980).

∗To whom correspondence should be addressed.

© The Author 2008. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org.
In addition to ascertainment of exposures related to the principal disease outcome of interest, researchers sometimes collect information on a secondary disease for each of the cases and controls in the study. Also, in countries where national disease registers have good coverage (e.g. in Scandinavia), information on the secondary outcome can be easily gathered later, even if the original study did not collect this information. If the risk factors for this secondary disease are also of interest, routine application of ordinary or conditional logistic regression will generally produce biased estimates of odds ratios. Recently published methodological work has demonstrated how to reuse data obtained from a case–control study for the analysis of such a secondary outcome. Reilly and others (2005) present a weighted likelihood approach, in which the contribution of each individual to the log-likelihood of the secondary outcome is weighted by the inverse of their probability of being selected into the study. Jiang and others (2006) propose several semiparametric approaches which involve explicit modeling of the association between the main and secondary outcome conditional on the covariates: one of these approaches reduces to Palmgren’s model for bivariate logistic regression (Palmgren, 1989). None of these approaches considers the nested case–control design. In a nested case–control study, incident cases of a disease that occur in a defined cohort are identified, and for each case a specified number of matched controls is selected from cohort members who are disease-free at the time of diagnosis of the case. The odds ratios from conditional logistic regression of the matched sets yield valid estimates of the hazard ratios that would have been obtained from a Cox regression analysis of the whole cohort (Prentice and Breslow, 1978). The nested case–control design potentially offers impressive reductions in the costs and efforts of data collection and analysis compared with the full cohort approach and is widely used in the study of chronic diseases.

The ability to reuse data is of growing importance. For instance, many epidemiological studies now gather expensive molecular and/or genetic data. With the advent of biobanking and follow-up of large population-based cohorts (e.g. UK biobank, http://www.ukbiobank.ac.uk/), investigations are no longer focused on a single primary outcome. While separate nested case–control studies could be conducted for each outcome of interest in such cohorts, this does not make optimal use of the investment in data collection. Our objective in this paper is to derive a method of reusing data from a previous nested case–control study to supplement a new nested case–control study whose main outcome of interest was a secondary outcome or covariate in the first study. The likelihood that is used for inference in a nested case–control study can be derived from a survival data perspective (see, e.g. Prentice and Breslow, 1978). This suggests that the use of models such as the copula model for multivariate survival data (Clayton, 1978) is needed if we are to model explicitly the association between the 2 outcomes. Here, we present an alternative approach based on weighted likelihood in which the extra modeling step is avoided. We consider a situation where the second study gathers only a few controls with the intention of supplementing this sparse information with existing data from the first.

2. Methodology

We first introduce some definitions associated with a study cohort and describe the 2 independent case–control studies nested within the cohort, before outlining our approach to combining the 2 studies. For simplicity, we assume that the hazard functions of the 2 diseases follow the Cox proportional hazards model:

\[
\lambda_{iA}(t) = \lambda_{0A}(t) \exp\{\beta'_A X_i + \gamma'_A Z_i(t)\}
\]

for disease A and

\[
\lambda_{iB}(t) = \lambda_{0B}(t) \exp\{\beta'_B X_i + \gamma'_B Z_i(t)\}
\]

for disease B, where \(X_i\) and \(Z_i(t)\) are matrices of fixed and time-dependent covariates for individual \(i\).
2.1 The cohort

Assume a cohort of \( N \) individuals. Individual \( i \) enters the cohort at starting time \( s_i \) and is followed up to exit at time \( e_i \). For this individual, the “onset” of disease \( A \) is at time \( T_{iA} \) and disease \( B \) at time \( T_{iB} \). The scope of the word “onset” here includes time of first diagnosis or time of first hospitalization and will depend on the data available. We assume that the 2 diseases are not competing risks, in the sense that the occurrence of one disease will not prevent the occurrence of the other. However, the hazard rate of a disease that has yet to occur is allowed to be altered after the occurrence of the other disease, through the use of the time-dependent covariates \( Z(t) \).

2.2 The first nested case–control study

The outcome of interest in the first study is the occurrence of disease \( A \). At the time of onset of disease \( A \) in individual \( i \) in the cohort, \( T_{iA} \) (where \( s_i < T_{iA} \leq e_i \)), \( m_1 \) matched controls are randomly selected from the risk set \( R_i \), that is, the set of all individuals with the same values of the matching variables as individual \( i \) but who are free of disease \( A \) at time \( T_{iA} \). We denote the subset of selected individuals from the risk set as \( R_i \). Valid estimates of \( \theta_A = (\beta_A, \gamma_A) \) can be obtained by maximizing the partial likelihood (Liddle and others, 1977)

\[
\ell(\theta_A) = \prod_{i \in D_A} \left[ \frac{\exp[\beta_A' X_i + \gamma_A' Z_i(T_{iA})]}{\sum_{k \in R_i} \exp[\beta_A' X_k + \gamma_A' Z_k(T_{iA})]} \right],
\]

(2.1)

where \( D_A \) indicates the set of cases of disease \( A \) that were identified during follow-up of the cohort. We assume that for each of the sampled individuals, in addition to information on fixed covariates and history of time-dependent covariates we also have information on the time of onset of disease \( B \).

2.3 The second nested case–control study

The outcome of interest in the second study is the occurrence of disease \( B \). At the time of onset \( T_{iB} \) of disease \( B \) in individual \( i \) in the cohort, \( m_2 \) matched controls (where \( m_2 < m_1 \)) are randomly selected from the risk set \( R_i \) of all individuals with the same matching variables as individual \( i \) but who are free of disease \( B \) at time \( T_{iB} \). Denoting the subset of individuals whose time of onset occurs before the end of their follow-up time as \( D_B \), valid estimates of \( \theta_B = (\beta_B, \gamma_B) \) can be obtained by maximizing the partial likelihood

\[
\ell(\theta_B) = \prod_{i \in D_B} \left[ \frac{\exp[\beta_B' X_i + \gamma_B' Z_i(T_{iB})]}{\sum_{k \in R_i} \exp[\beta_B' X_k + \gamma_B' Z_k(T_{iB})]} \right].
\]

(2.2)

In addition to information on fixed covariates and the history of time-dependent covariates, we assume that information is available on the time of onset of disease \( A \) for each of the sampled individuals.

2.4 Combining 2 studies to analyze risk factors for the second outcome

When there are only 1 or 2 controls for each case in the second study, then the estimates obtained by maximizing \( (2.2) \) will not be very precise. If there are sufficient resources, then efficiency can be increased by sampling additional controls for each case. However, we propose instead to reuse data already collected in the first study to improve the efficiency of the second. If there is no association between the 2 diseases, then with respect to disease \( B \), data from the first study can be thought as a random subset of the original cohort (conditional on matching covariates), and hence valid estimates of \( \theta_B = (\beta_B, \gamma_B) \) can be obtained.
Combining nested case–control data

from the first data set by maximizing the following partial likelihood:

\[
\ell(\theta_B) = \prod_{i \in D_{BA}} \left[ \frac{\exp[\beta_X' X_i + \gamma_Z' Z_i(T_iB)]}{\sum_{k \in R_i} \exp[\beta_X' X_k + \gamma_Z' Z_k(T_iB)]} \right],
\]

(2.3)

where \(D_{BA}\) is the subset of individuals in the first data set whose onset of disease \(B\) occurs before the end of their follow-up time and the summation in the denominator covers all matching individuals in the first data set who have not yet experienced event \(B\) at time \(T_iB\). The estimates from the combined data set can be obtained by maximizing the product of the likelihood functions in (2.2) and (2.3).

When there is an association between the 2 diseases, the simple approach above will yield biased estimates of \(\beta_B\) and \(\gamma_B\). This is because the first study sampled all cases of disease \(A\), while only a fraction of individuals who have not experienced \(A\) are sampled at each time point when a case occurs. In a situation where the 2 diseases are positively associated, in the sense that those who experience disease \(A\) are more likely to also experience disease \(B\), the sampled individuals in the first study will overrepresent those who experienced disease \(B\) while at the same time underrepresent those who did not. This suggests that a Horvitz–Thompson approach, with appropriate weights, may enable us to obtain unbiased estimates of \(\beta_B\) and \(\gamma_B\).

The Horvitz–Thompson approach weights the prospective log-likelihood of each individual by the inverse of the probability that they are selected into the sample. In our case, we wish to combine the 2 data sets using appropriate weights. Let us denote as \(\Omega_A\) the set of individuals selected in the first study and as \(\Omega_B\) the set of individuals selected in the second study. Our objective is to estimate \(\theta_B = (\beta_B, \gamma_B)\) based on data from all individuals in the 2 studies, that is, based on \(\Omega_i = \Omega_A \cup \Omega_B\). Applying the Horvitz–Thompson approach to our problem, we will estimate \(\beta_B, \gamma_B\) by maximizing the following weighted log-likelihood:

\[
\log \ell_{w}(\theta_B) = \sum_{i \in \Omega_i} w_i \log \ell_i(\theta_B|T_iB, s_i, e_i, X_i, Z_i(t)),
\]

(2.4)

where \(w_i = 1/p_i\), \(p_i\) being the probability that individual \(i\) is selected in either study. The component \(\log \ell_i(\theta_B|T_iB, s_i, e_i, X_i, Z_i(t))\) is the log-likelihood of individual \(i\) experiencing disease \(B\) at time \(T_iB\) given their fixed covariates and history of time-dependent covariates. If the disease occurred before the end of follow-up, then the log-likelihood contribution is given by

\[
\log \lambda_iB(T_iB) - \int_{s_i}^{T_iB} \dot{\lambda_iB}(t) dt,
\]

and if no onset has occurred by the end of follow-up, the log-likelihood is given by

\[- \int_{s_i}^{e_i} \dot{\lambda_iB}(t) dt.\]

The probability \(p_i\) of individual \(i\) being selected into either study can be written as

\[
p_i = p_{iA} + p_{iB} - p_{iA} p_{iB},
\]

(2.5)

where \(p_{iA}\) and \(p_{iB}\) are the probabilities that individual \(i\) is selected as part of study \(A\) and study \(B\), respectively. Since nested case–control studies always sample individuals with disease onset, the probability of
being selected in either study is 1 for those who develop at least one of the 2 diseases before the end of follow-up. For those who do not experience either event before the end of follow-up, the expression for their probability of being selected is more complicated. The probabilities of individual $i$ not being selected can be written as

$$1 - p_{iA} = \prod_{j \in \Omega_A, s_i \leq T_{jA} \leq T_i} \left[ 1 - \frac{m_1}{Y_j^A - 1} I(U_j = U_i) \right],$$

$$1 - p_{iB} = \prod_{j \in \Omega_B, s_i \leq T_{jB} \leq T_i} \left[ 1 - \frac{m_2}{Y_j^B - 1} I(V_j = V_i) \right],$$

where $T_{iA}$ and $T_{iB}$ are the follow-up times (to disease or censoring) of individual $i$ in studies A and B respectively, $Y_j^A$ is the number of individuals in the cohort still at risk for disease A at time $T_{jA}$, and $Y_j^B$ is the number of individuals in the cohort still at risk for disease B at time $T_{jB}$. $U_i$ denotes the values of the matching variables of individual $i$ in the first study, and $I(U_j = U_i)$ is an indicator function for whether individual $i$ has the same values of matching variables as the individual $j$ who experienced onset of disease A. Likewise, $I(V_j = V_i)$ is an indicator function that applies to the second study and is to one if individual $i$ has the same values of matching variables as the individual $j$ who experienced onset of disease B. Hence, for individual $i$ the multiplication behind the product signs will only take place each time an individual with the same matching variables as individual $i$ has onset while individual $i$ is still disease-free and thus can be considered as a candidate for the set of controls.

The resulting score equation of the log-likelihood in (2.4) takes a similar form to the estimating equation in the generalized estimating equation (GEE) approach for correlated data. As is the case with GEE estimates, the estimates we get from maximizing (2.4) are unbiased and their variances are given by the information matrix of (2.6). Although

$$\Delta \approx \sum_{i \in \Omega_i} \left( 1 - \frac{p_i}{p_i^2} \right) S_i'(\theta_B) S_i(\theta_B) + \sum_{i \in \Omega_i} \sum_{i' \neq i} \left( \frac{q_{i,i'} - (1 - p_i)(1 - p_{i'})}{p_i^2 p_{i'}^2} \right) S_i'(\theta_B) S_{i'}(\theta_B),$$

where $S_i(\theta_B) = \frac{\partial \log f_i(\theta_B)}{\partial \theta_B}$ is the unweighted score vector for individual $i$ and $q_{i,i'}$ is the probability that individuals $i$ and $i'$ are not included in either study.

### 2.5 Baseline hazard assumption and proportional hazards models

The weighted likelihood approach in (2.4) requires parametric specification of the baseline hazard function. Given that misspecification of the baseline hazard will lead to biased and inconsistent estimates of the covariate effects, it is desirable to avoid the necessity of specifying the function. For proportional hazards models, this is possible. Samuelsen (1997) developed a semiparametric pseudolikelihood approach for nested case–control data whereby the contribution of controls to the denominator in the partial likelihood is weighted by the inverse of their probability of inclusion. We adapt this to our application, where we use a proportional hazards model. We obtain the estimates from this semiparametric approach using routine Cox proportional hazards regression with the sampling weight for individual $i$ given by (2.5). Although
the standard error of the estimates will not be correct, these estimates are useful as benchmark values and we use them for informal checking of the proportional hazards assumption by plotting the Schoenfeld residuals against time (Therneau and Grambsch, 2000, Chapter 6).

3. Application

We demonstrate our methods using 2 nested case–control studies from a simulated cohort (see supplementary material available at Biostatistics online, http://www.biostatistics.oxfordjournals.org) and then analyze data from a study of risk factors for anorexia in a cohort of young women in Sweden (Cnattingius and others, 1999) combined with a study of risk factors for schizophrenia and affective and reactive psychosis (Hultman and others, 1999) in an overlapping cohort. We compare the risk estimates obtained using (a) data from the second study only and (b) the combined data from both studies, obtained by maximizing expressions (2) and (4), respectively.

3.1 Anorexia nervosa study

Cnattingius and others (1999) investigated the association between several birth factors and the risk of anorexia nervosa using a nested case–control study of girls born in Sweden between 1973 and 1984. Using the Swedish inpatient register, they extracted the date of first diagnosis for 781 young women from all over Sweden who were diagnosed with anorexia nervosa between 10 and 21 years of age. For each of these cases, they selected 5 controls, individually matched on the year of birth and hospital of birth. This choice of the number of controls is common in case–control studies, as additional controls beyond 4 or 5 per case will not substantially increase the relative efficiency of the study compared to a full cohort study (see, e.g. Walter, 1980). Although the number of controls per case in the anorexia study is adequate, for the purpose of our demonstration here we will disregard 4 of the 5 selected controls to create an example with only one matched control per case. To increase the precision of estimates from this “new” data set, we combine it with data from an earlier nested case–control study of schizophrenia, affective psychosis, and reactive psychosis (Hultman and others, 1999). These data were collected to study prenatal and perinatal risk factors for these conditions in children born in Sweden between 1973 and 1979, and we will refer to this second data set as the Schizophrenia, Affective and Reactive Psychosis (SARP) data set. In this study, 657 children were diagnosed with any one of the 3 diseases of interest. For each case, 5 controls were selected, individually matched on year of birth and hospital of birth. Although the original data set used for this study did not have information on the date of onset of anorexia, this information can be easily obtained from the Swedish hospital discharge register, using the unique national registration numbers of the individuals. Our interest is in using the additional information available from the follow-up of the girls in the schizophrenia study to supplement the information in the anorexia data. In the schizophrenia study, the follow-up times ranged from 14.96 to 21.96 years with an average of 18.23 years, compared to a range from 10.28 to 21.31 years and an average of 15.30 years in the anorexia study.

For brevity, we focus our attention on covariates found to be significant risk factors by Cnattingius and others (1999): indicator variables for gestational age less than 33 weeks and “birth trauma” defined by International Statistical Classification of Diseases and Related Health Problems 8 code 772. Since both studies used hospital and year of birth as matching variables, the annual counts of newborn baby girls in the different hospitals are needed to compute the appropriate weights for combining the 2 data sets. These counts were obtained from the Swedish medical birth register.

Table 1 presents the estimates of the effects of the binary risk factors obtained by using the anorexia data set only and the combined data set. Proportional hazards models are assumed in both analyses. Since we need to specify the form of the baseline hazard function, we present estimates obtained under
Table 1. Log hazard ratio estimates with anorexia as outcome variable

<table>
<thead>
<tr>
<th>Data set</th>
<th>(a) Gestational age &lt; 33 weeks</th>
<th>(b) Birth trauma</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimates</td>
<td>SE</td>
</tr>
<tr>
<td>Anorexia</td>
<td>1.266</td>
<td>0.570</td>
</tr>
<tr>
<td>Anorexia + SARP, exponential baseline hazard</td>
<td>0.787</td>
<td>0.371</td>
</tr>
<tr>
<td>Anorexia + SARP, Weibull baseline hazard</td>
<td>0.770</td>
<td>0.376</td>
</tr>
</tbody>
</table>

Fig. 1. Plots of smoothed Schoenfeld residuals for checking the proportional hazards assumption. (a) Gestational age. (b) Birth trauma.

assumptions of exponential and Weibull hazard functions. The estimates for the combined data sets are very similar under the 2 different assumptions regarding the baseline hazard functions. For the anorexia data, with only one control per case, although we could detect the significance of both risk factors the standard errors of estimates are rather large. However, on combining the 2 data sets we obtained estimates with lower standard errors and thus increased the power of detecting the significance of the risk factors. The relative efficiency gain is approximately 55% for gestational age and 41% for birth trauma, although the gain for gestational age needs to be interpreted cautiously since there is a reduction of effect size in the combined analysis.

To check the proportional hazards assumption, we fitted a “weighted” Cox proportional hazards model with sampling weights for individuals given by their probability of being included in at least one of the 2 studies. The plots of the Schoenfeld residuals against time for gestational age and birth trauma are shown in Figure 1(a) and (b), respectively. No residual trend is observed in either variable. Furthermore, the estimates of gestational age and birth trauma effects from the weighted Cox proportional hazard of the combined data are 0.775 and 0.462, respectively, very similar to the estimates in Table 1 for the Weibull
baseline hazard function. Hence, we can safely assume that the proportional hazards assumption with Weibull baseline hazard is reasonable for our data.

4. DISCUSSION

We have demonstrated the feasibility of reusing data already collected in a nested case–control study to supplement another nested case–control study of the same or overlapping cohort with a different main outcome variable. In an attempt to use all available controls, we “break” the matching and use a prospective weighted likelihood approach, whereby the contribution of an individual to the prospective log-likelihood is weighted by the inverse of the probability of the individual being included in either study (2.4). We illustrated the method with an application to a study of anorexia and demonstrated the efficiency that could be gained from supplementing this study with data from a study of schizophrenia in an overlapping cohort.

Our approach requires us to specify the baseline hazard function and, in the case of time-dependent covariates, to have complete history of any important covariates in order to enable the computation of the cumulative hazard. Because of the need to specify the baseline hazard, the method may be seen as inferior to the semiparametric partial likelihood approach for Cox proportional hazards models. On the other hand, modeling the baseline hazard allows more flexibility when the proportional hazards assumption is violated (for testing proportional hazards assumption, see, e.g. Therneau and Grambsch, 2000, Chapter 6), in which case alternative models such as accelerated lifetime or additive hazard models may be appropriate. To guard against serious consequences of baseline hazard misspecification, we suggest that users fit a weighted Cox proportional hazards model with appropriate sampling weights for individuals given by their probability of being included in either study. The estimates from this model can be used to conduct an informal check of the proportional hazard assumption; a plot of the Schoenfeld residuals against time will highlight trends that indicate nonproportionality. Even if no significant trend is observed but the estimates of fixed covariate effects differ markedly from the estimates from a weighted Cox regression, this suggests that the baseline hazard function is not appropriate. Regarding the requirement that a complete history of time-dependent covariates is available, this is only necessary if we believe that the hazard function is modified by these covariates. Most likely, such covariates will be available in the form of repeated measurements on the same individual at discrete time points, in which case we can approximate the complete history by assuming the covariates to be piecewise constant in time.

The gain in terms of relative efficiency when reusing data from another nested case–control study will depend on 2 main factors: the average follow-up time in the first study, with respect to the outcome in the second, and the probability that individuals in the first study are included in the combined study. The form of the asymptotic variance estimates in (2.6) suggests that the penalty term is larger for individuals with smaller probability of inclusion. Thus, cases from the first study (for whom \( w_i = 1 \)) contribute information as do any other individuals with large probability of inclusion. In a large cohort, the controls in the first study have only a small probability of inclusion, except they develop the disease of interest, but in this case they would have appeared anyway as cases in the second study. In practical terms, this means that the efficiency gain depends on the number of cases with disease \( A \) relative to the number of cases with disease \( B \), in the combined data set. The actual efficiency gain will also be affected by the relative average follow-up time in the first and second study, with respect to the second outcome. Thus, if the first study has much fewer cases and significantly shorter length of follow-up (with respect to the second outcome), then a full joint analysis would not be expected to result in meaningful efficiency gain. In both of our illustrations, application of the methodology resulted in considerable gain in efficiency from the joint analysis. In the simulated example, the efficiency gain is driven by the fact that there are approximately 125 cases of disease \( A \) in the first study as compared to only 50 cases of disease \( B \) in
the second. In the anorexia example, there are only 274 girls with onset of SARP as compared to 781 girls with onset of anorexia in the current study, but the efficiency is boosted by the fact that the average follow-up time in the SARP study (18.23 years) is longer than that in anorexia study (15.30 years).

In conclusion, our method enables researchers planning nested case–control studies to consider reusing data that may have been gathered for previous studies in the same or overlapping cohort. To be appropriate for reuse, the data from the first study must include not only the outcome of interest in the second study but also the covariates. In many studies of cancer epidemiology, such data are likely to be available, since outcome information is often taken from national cancer registers (which record all cancers), and many of the covariates that are usually collected are risk factors that are common to several cancers. For diseases other than cancer, appropriate outcome data will also be available in countries where hospital and disease registers have good coverage, such as in Scandinavia. Information on the onset of the second outcome can be easily extracted from such registers when needed, even though the original study did not collect this information. The potential risk factors for the second outcome may pose more of a problem, except in applications where the first and second outcomes share similar pathways, so that important risk factors for the second outcome would likely be collected in the first study. It is also possible to combine studies that have used different matching variables, provided the individual values of matching variables used in one study but not in the other are available to enable computation of the inclusion probability. Since matching variables are usually important demographic characteristics, these are likely to be measured in various studies in the same cohort. While our method can be useful in traditional population-based epidemiology studies, it may be more crucial to investigations in genetic and molecular epidemiology, where nested case–control studies use expensive exposure measurements from specimens sampled from large biobanks.

**FUNDING**

Australian National Health and Medical Research Council Grant (418020 to A.S.).

**ACKNOWLEDGMENTS**

We would like to thank Anna Svensson for extracting the counts of births by year and hospital from the Swedish medical birth register, and Andrew Mackinnon for perusing the early draft of this manuscript. *Conflict of Interest*: None declared.

**REFERENCES**


*Received July 23, 2007; first revision March 11, 2008; second revision April 7, 2008; accepted for publication May 12, 2008*