Biostatistics (2009), **10**, 2, *pp.* 364–373
doi:10.1093/biostatistics/kxn043
Advance Access publication on February 6, 2009

A robust method for finely stratified familial studies with proband-based sampling

MOLIN WANG*

Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute,
44 Binney Street, Boston, MA 02115, USA
mwang@jimmy.harvard.edu

JOHN J. HANFELT

Department of Biostatistics, Rollins School of Public Health, Emory University,
Atlanta, GA 30322, USA

SUMMARY

This paper presents a robust method to conduct inference in finely stratified familial studies under proband-based sampling. We assume that the interest is in both the marginal effects of subject-specific covariates on a binary response and the familial aggregation of the response, as quantified by intrafamilial pairwise odds ratios. We adopt an estimating function for proband-based family studies originally developed by Zhao *and others* (1998) in the context of an unstratified design and treat the stratification effects as fixed nuisance parameters. Our method requires modeling only the first 2 joint moments of the observations and reduces by 2 orders of magnitude the bias induced by fitting the stratum-specific nuisance parameters. An analytical standard error estimator for the proposed estimator is also provided. The proposed approach is applied to a matched case–control familial study of sleep apnea. A simulation study confirms the usefulness of the approach.

Keywords: 2-Index asymptotics; Adjusted profile estimating function; Ascertainment bias; Bias reduction; Familial aggregation; Nuisance parameter; Proband; Sparse data; Stratified study.

1. INTRODUCTION

Familial aggregation studies are often conducted by genetic epidemiologists to investigate both the genetic and the environmental causes of disease. In such studies, the aim is usually to assess the familial aggregation of the response as well as the marginal effects of subject-specific covariates on the response. Special features of such studies, such as the presence of ascertainment bias, the need to stratify the design to control for heterogeneity of risk of disease in the population, and the fact that the second joint moments of clustered data are of interest, can make inference complicated.

Proband-based sampling (e.g. Neuhaus *and others*, 2002) is a commonly used design to ascertain families in the study of a rare disease. Under this study design, families are entered based on the known disease

*To whom correspondence should be addressed.

status of a single family member (i.e. proband), with subsequent recruitment of the family members of each proband. We make the popular and convenient "single-ascertainment" assumption, which posits that a suitable method to adjust for ascertainment bias of families is to regard the probands' disease statuses as fixed (Williamson *and others*, 1996; Zhao *and others*, 1998). Our focus is on responses that are binary, in which case the single-ascertainment assumption gives rise to the case–control family design in which relatives of case probands are compared to relatives of control probands. As noted by Liang and Beaty (2000), fine stratification of a family case–control study is desirable when family members share similar values of the stratification variables. Our chief interest is in the following application of a finely stratified familial aggregation study with proband-based sampling.

Example: sleep apnea study. In a study of sleep disturbance, patients with obstructive sleep apnea (OSA) diagnosed by measurement of the respiratory disturbance index (RDI) and a medical history consistent with the disorder were recruited as case probands (Williamson *and others*, 1996). Each proband provided a list of 3 neighbors, one of whom was randomly selected as the matched control. All available first-degree relatives were invited to participate in the study. Matching families by neighborhood of residence served to control, at least partially, for the potentially confounding effects of socioeconomic status and environment. A previous analysis of this study by Wang *and others* (2004) was based on a continuous response, RDI. Here, the focus is on an arguably more clinically relevant outcome, presence of OSA. Since the medical history was not available for the relatives, the outcomes of relatives were defined only by the RDI measurements (presence of OSA if RDI $\geqslant 15$, absence otherwise; Lowe *and others*, 2000). The scientific interest was in the effects of key covariates such as age, body mass, and gender on risk of OSA as well as the intrafamilial pairwise association, quantified by odds ratio of presence of OSA in each pair of family members.

Several methods are available for the analysis of family data with a binary outcome under proband-based sampling in the special case of an unstratified study design. Whittemore (1995) applies a multivariate logistic regression model and estimates the correlation among family members and the relationship between covariates and disease based on maximum likelihood. Zhao *and others* (1998) present a robust estimating equation approach, which requires modeling only the first 2 joint moments of the observations, to estimate consistently both the marginal effects of covariates and the intrafamilial pairwise correlation coefficient. These 2 approaches may not apply in finely stratified family studies because the estimation may be highly biased in the situation where there is a large number of stratum-specific nuisance parameters and the number of families in each stratum is relatively small; see Section 2.3 below.

Few methods have been proposed for finely stratified family studies with a binary outcome and proband sampling. Liang and Pulver (1996) applied a type of conditional generalized estimating equation model to analyze family data under either a matched or an unmatched proband sampling design, where the probands' outcomes are included as covariates in the regression modeling. This approach does not allow one to assess the marginal risk factors for disease. Hanfelt's (2004) composite conditional likelihood approach provides inferences about the effects of covariates on marginal response probabilities and the pairwise odds ratio of family members in the context of a finely stratified study; however, asymptotic properties of the estimator are not known fully.

In this paper, we consider an estimating function originally developed by Zhao *and others* (1998) in the context of an unstratified study design but include an additive bias reduction term, based on the adjusted profile estimating function approach (Wang and Hanfelt, 2008), without which the bias induced by the fitting of stratum-specific nuisance parameters would be severe under a finely stratified design. Knowledge of only the first 2 joint moments is required to form the estimating function and its adjustment term. Asymptotic properties of our estimator are provided.

The proposed method requires at least one affected relative of the probands in each stratum and is intended for applications with reasonably high disease rates; see Section 6 below.

This paper is organized as follows. Section 2 presents background on the estimating functions for the main effects and the intrafamilial pairwise odds ratios in the context of a study design that is not finely stratified. In Section 3, we return the focus to a finely stratified design and discuss an adjustment approach to reduce the bias of the profile estimating function. A simulation study is presented in Section 4 and the supplementary material available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org). An analysis of the matched case–control familial study of sleep apnea appears in Section 5. We conclude with a brief discussion.

2. ESTIMATING FUNCTIONS FOR FAMILY STUDIES WITH PROBAND-BASED SAMPLING

2.1 *Models and assumptions*

Let d_{ijk} and $\mathbf{x}_{ijk}$ denote a binary outcome variable and a vector of covariates for the kth subject in the jth family of the ith stratum, $k = 0, \ldots, m_{ij}$; $j = 1, \ldots, n_i$; $i = 1, \ldots, N$. Here, $k = 0$ refers to the proband. Denote the stratification variable in the ith stratum as u_i. As in Zhao *and others* (1998), we adopt a marginal univariate logistic regression model under the subject-specific effect assumption, so that the association between an individual's outcome d_{ijk} and the family's collection of covariates is given by

$$\begin{aligned}\mu_{ijk} &= \mathrm{pr}\left(d_{ijk} = 1 \left|\mathbf{x}_{ij0,}\mathbf{x}_{ij1}, \ldots, \mathbf{x}_{ijm_{ij}}, u_i\right.\right) = \mathrm{pr}(d_{ijk} = 1|\mathbf{x}_{ijk}, u_i)\\ &= \{1 + \exp(-\alpha_{0i} - \alpha'\mathbf{x}_{ijk})\}^{-1}, \quad k = 0, \ldots, m_{ij}, \end{aligned} \tag{2.1}$$

where $\alpha = (\alpha_1, \ldots, \alpha_l)$ is a vector of parameters of interest and the stratum-specific intercepts α_{0i}, which are the effects of the stratification variable, are considered as nuisance parameters.

As an alternative to the intrafamilial correlation coefficient in Zhao *and others* (1998), we use the intrafamilial pairwise odds ratio,

$$r_{ijk_1k_2} = \frac{\mathrm{pr}\big(d_{ijk_1} = 1, d_{ijk_2} = 1\big)\mathrm{pr}\big(d_{ijk_1} = 0, d_{ijk_2} = 0\big)}{\mathrm{pr}\big(d_{ijk_1} = 1, d_{ijk_2} = 0\big)\mathrm{pr}\big(d_{ijk_1} = 0, d_{ijk_2} = 1\big)},$$

to quantify the familial aggregation. We model the odds ratio using regression with a given link function $h(\cdot)$ to give

$$r_{ijk_1k_2} = h\left(\beta^{\mathrm{T}}\mathbf{z}_{ijk_1k_2}\right), \quad 0 \leqslant k_1 < k_2 \leqslant m_{ij}, \tag{2.2}$$

where $\beta = (\beta_1, \ldots, \beta_p)$ is a vector of parameters of interest and $\mathbf{z}_{ijk_1k_2}$ is a vector of covariates. For example, $\mathbf{z}_{ijk_1k_2}$ might include an indicator of the genealogical relationship between the k_1th and k_2th members of the ijth family. An intercept term is allowed in the odds ratio regression model (2.2). Hence, if the odds ratio depends on the stratification variable, we can change the intercept β_1 in (2.2) to be a stratum-specific nuisance parameter, say β_{1i}, and the proposed method can easily be extended to make inferences on $(\beta_2, \ldots, \beta_p)$.

2.2 *Estimating functions*

Under proband-based sampling, the marginal mean of a proband's outcome in the ith stratum can be modeled as (Whittemore, 1995; Zhao *and others*, 1998)

$$\upsilon_{ij0} = \mathrm{pr}(d_{ij0} = 1|\mathbf{x}_{ij0}, u_i; \lambda_i, \alpha) = \{1 + \exp(\lambda_i - \alpha'\mathbf{x}_{ij0})\}^{-1},$$

where $\lambda_i = \log\{(1 - \pi_i)\eta_i\}/\{\pi_i(1 - \eta_i)\} - \alpha_{0i}$, η_i is the disease prevalence for the general population corresponding to the ith stratum, π_i is the proportion of cases among the probands in the ith stratum, and the λ_i are treated as nuisance parameters.

Following the approach of Zhao *and others* (1998) and Williamson *and others* (1996), we adjust for ascertainment bias by conditioning on the probands' disease statuses, which can be justified by the single-ascertainment assumption. The relevant conditional probability $\mathrm{pr}(d_{ijk} = 1 | d_{ij0}, \mathbf{x}_{ij0}, \mathbf{x}_{ijk}; \alpha_{0i}, \alpha, \beta)$, denoted as $\upsilon_{ijk}(\alpha_{0i}, \alpha, \beta)$, can be evaluated using Bayes rule; see Section A of the supplementary material available at *Biostatistics* online.

An estimating function for α and β, which depends on the stratum-specific nuisance parameters λ_i and α_{0i}, $i = 1, \ldots, N$, is given by $\mathbf{s} = (\mathbf{s}_1^{\mathrm{T}} + \mathbf{s}_2^{\mathrm{T}}, \mathbf{s}_3^{\mathrm{T}})^{\mathrm{T}}$, where

$$\mathbf{s}_1(\alpha; \lambda_1, \ldots, \lambda_N) = \sum_{i=1}^{N} \mathbf{s}_{1i}(\alpha; \lambda_i) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{s}_{1ij}(\alpha; \lambda_i) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left(\frac{\partial \upsilon_{ij0}}{\partial \alpha} \right)^{\mathrm{T}} \frac{d_{ij0} - \upsilon_{ij0}}{\upsilon_{ij0}(1 - \upsilon_{ij0})}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{x}_{ij0}(d_{ij0} - \upsilon_{ij0}),$$

$$\mathbf{s}_2(\alpha, \beta; \alpha_{01}, \ldots, \alpha_{0N}) = \sum_{i=1}^{N} \mathbf{s}_{2i}(\alpha, \beta; \alpha_{0i}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{s}_{2ij}(\alpha, \beta; \alpha_{0i})$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left(\frac{\partial \mathbf{v}_{ij}}{\partial \alpha} \right)^{\mathrm{T}} V_{22ij}^{-1}(\mathbf{d}_{ij} - \mathbf{v}_{ij}) \tag{2.3}$$

and

$$\mathbf{s}_3(\alpha, \beta; \alpha_{01}, \ldots, \alpha_{0N}) = \sum_{i=1}^{N} \mathbf{s}_{3i}(\alpha, \beta; \alpha_{0i}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{s}_{3ij}(\alpha, \beta; \alpha_{0i})$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left(\frac{\partial \mathbf{v}_{ij}}{\partial \beta} \right)^{\mathrm{T}} V_{22ij}^{-1}(\mathbf{d}_{ij} - \mathbf{v}_{ij}). \tag{2.4}$$

In the above, $\mathbf{d}_{ij}^{\mathrm{T}} = \{d_{ij1}, \ldots, d_{ijm_{ij}}\}$, $\mathbf{v}_{ij}^{\mathrm{T}} = \{\upsilon_{ij1}, \ldots, \upsilon_{ijm_{ij}}\}$, $V_{22ij} = A_{ij}^{1/2} W_{ij} A_{ij}^{1/2}$, A_{ij} is an $m_{ij} \times m_{ij}$ diagonal matrix with kth diagonal entry $\upsilon_{ijk}(1 - \upsilon_{ijk})$, and W_{ij} is the correlation matrix of $\mathbf{d}_{ij}$ conditional on d_{ij0}. Specification of the true conditional correlation matrix W_{ij} would require additional modeling assumptions about the intrafamiliar 3-way associations of the responses.

Estimating functions for nuisance parameters λ_i and α_{0i} are, respectively,

$$h_{1i}(\lambda_i; \alpha) = \sum_{j=1}^{n_i} h_{1ij}(\lambda_i; \alpha) = \sum_{j=1}^{n_i} \frac{\partial \upsilon_{ij0}}{\partial \lambda_i} \frac{d_{ij0} - \upsilon_{ij0}}{\upsilon_{ij0}(1 - \upsilon_{ij0})} = -\sum_{j=1}^{n_i} (d_{ij0} - \upsilon_{ij0}),$$

$$h_{2i}(\alpha_{0i}; \alpha, \beta) = \sum_{j=1}^{n_i} h_{2ij}(\alpha_{0i}; \alpha, \beta) = \sum_{j=1}^{n_i} \left(\frac{\partial \mathbf{v}_{ij}}{\partial \alpha_{0i}} \right)^{\mathrm{T}} V_{22ij}^{-1}(\mathbf{d}_{ij} - \mathbf{v}_{ij}),$$

for $i = 1, \ldots, N$. See Section B of the supplementary material available at *Biostatistics* online for derivation of the estimating functions.

2.3 *Limitations of the estimating functions*

Note that $\mathbf{s}_{2i} = \sum_j \mathbf{s}_{2ij}$ is a sum of n_i independent estimating functions; moreover, $\mathbf{s}_{2ij}$ is an unbiased estimating function in the sense that $E_{\mathbf{d}_{ij},\mathbf{x}_{ij}|\mathbf{d}_{i0}}(\mathbf{s}_{2ij}|\mathbf{d}_{i0}) = E_{\mathbf{x}_{ij}|\mathbf{d}_{i0}}\{E_{\mathbf{d}_{ij}|\mathbf{x}_{ij},\mathbf{d}_{i0}}(\mathbf{s}_{2ij}|\mathbf{x}_{ij},\mathbf{d}_{i0})\} = 0$, where $\mathbf{d}_{i0} = (d_{i10},\ldots,d_{in_i0})$ and $\mathbf{x}_{ij}$ is a collection of the ijth family's covariates $\mathbf{x}_{ij0},\ldots,\mathbf{x}_{ijm_{ij}}$. By similar arguments, $\mathbf{s}_{3i}$ and h_{2i} are also sums of n_i independent unbiased estimating functions, as are the proband-based estimating functions $\mathbf{s}_{1i}$ and h_{1i}. This may be seen by re-expressing $\mathbf{s}_{1i}$ and h_{1i} as $\mathbf{s}_{1i} = \sum_j \mathbf{s}_{1ij} = \sum_j \{\mathbf{s}_{1ij} - E(\mathbf{s}_{1ij}|d_{ij0})\}$ and $h_{1i} = \sum_j \{h_{1ij} - E(h_{1ij}|d_{ij0})\}$, respectively. We note that for an individual proband, the estimating functions $\mathbf{s}_{1ij}$ and h_{1ij} can have nonzero conditional expectation.

Inferential complications arise in finely stratified studies as a result of a large number of stratum-specific nuisance parameters, the fitting of which introduces non-negligible bias in the resulting profile estimating functions for the interest parameters α and β; see Section C of the supplementary material available at *Biostatistics* online for discussion. In the next section, we apply a method of adjusting profile estimating functions (Wang and Hanfelt, 2008) to reduce the sensitivity of the estimating functions $\mathbf{s}_2(\alpha,\beta;\alpha_{01},\ldots,\alpha_{iN})$, $\mathbf{s}_3(\alpha,\beta;\alpha_{01},\ldots,\alpha_{iN})$ to the stratum-specific nuisance parameters, $\alpha_{01},\ldots,\alpha_{iN}$. We do not apply the adjustment method for $\mathbf{s}_1(\alpha;\lambda_1,\ldots,\lambda_N) = \sum_i\sum_j\{\mathbf{s}_{1ij} - E(\mathbf{s}_{1ij}|d_{ij0})\}$ because the adjustment term for $\mathbf{s}_1$ would depend on $E_{\mathbf{x}_{ij0}|\mathbf{d}_{ij0}}(\mathbf{s}_{1ij}|d_{ij0})$ and $E_{\mathbf{x}_{ij0}|\mathbf{d}_{ij0}}(h_{1ij}|d_{ij0})$, specification of which requires assumptions about the probability model of probands' covariates.

3. ADJUSTED PROFILE ESTIMATING FUNCTIONS FOR STRATIFIED FAMILIAL STUDIES WITH PROBAND-BASED SAMPLING

3.1 *Adjusted profile estimating functions*

We now consider methods for adjusting each stratum-specific component $\mathbf{s}_{2i}, \mathbf{s}_{3i}$ of $\mathbf{s}_2, \mathbf{s}_3$, so that the adjusted profile estimating functions have lower order of bias than the naive profile estimating functions. We consider the robust method proposed by Wang and Hanfelt (2008), which can adjust a profile estimating function under fewer modeling assumptions. See Section D of the supplementary material available at *Biostatistics* online for discussion on why the adjustment methods in Wang and Hanfelt (2003) and Severini (2002) do not apply.

To take into account the proband-based sampling design in the familial studies considered in this paper, we use the conditional expectations $E(\cdot|\mathbf{d}_{i0})$ to replace all the marginal expectations in the Wang and Hanfelt (2008) method. Define

$$\mathbf{g}_{bi}(\alpha,\beta,\alpha_{0i}) = \mathbf{s}_{bi} - E\left(\frac{\partial \mathbf{s}_{bi}}{\partial\alpha_{0i}}\middle|\mathbf{d}_{i0}\right)E^{-1}\left(\frac{\partial h_{2i}}{\partial\alpha_{0i}}\middle|\mathbf{d}_{i0}\right)h_{2i}, \quad b = 2,3.$$

This satisfies an orthogonality property $E(\partial\mathbf{g}_{bi}/\partial\alpha_{0i}|\mathbf{d}_{i0}) = 0$, that is an important first step in achieving inference insensitive to the effects of fitting stratum-specific nuisance parameter α_{0i} (Wang and Hanfelt, 2008). Note that $\mathbf{g}_{bi}(\alpha,\beta,\hat\alpha_{0i(\alpha,\beta)}) = \mathbf{s}_{bi}(\alpha,\beta,\hat\alpha_{0i(\alpha,\beta)})$, where $\hat\alpha_{0i(\alpha,\beta)}$ is obtained by solving $h_{2i} = 0$.

The effect of fitting the nuisance parameter on $\mathbf{g}_{bi}(\alpha,\beta,\alpha_{0i})$ is approximately (Wang and Hanfelt, 2008)

$$\begin{aligned}
&\mathbf{g}_{bi}(\alpha,\beta,\alpha_{0i}) - \mathbf{g}_{bi}(\alpha,\beta,\hat\alpha_{0i(\alpha,\beta)})\\
&\quad = E^{-1}\left(\frac{\partial h_{2i}}{\partial\alpha_{0i}}\middle|\mathbf{d}_{i0}\right)E\left(h_{2i}\frac{\partial\mathbf{g}_{bi}}{\partial\alpha_{0i}}\middle|\mathbf{d}_{i0}\right)\Bigg|_{\alpha_{0i}=\hat\alpha_{0i(\alpha,\beta)}}\\
&\qquad -\frac{1}{2}E^{-2}\left(\frac{\partial h_{2i}}{\partial\alpha_{0i}}\middle|\mathbf{d}_{i0}\right)E\left(\frac{\partial^2\mathbf{g}_{bi}}{\partial\alpha_{0i}^2}\middle|\mathbf{d}_{i0}\right)E(h_{2i}^2|\mathbf{d}_{i0})\Bigg|_{\alpha_{0i}=\hat\alpha_{0i(\alpha,\beta)}} + z_{bi1} + z_{bi2}n_i^{-1/2} + \mathrm{O}_p(n_i^{-1}),
\end{aligned}$$

where z_{bi1} and z_{bi2} are mean-zero random variables. This leads to an adjusted profile estimating function for a single stratum i that reduces the bias by 2 orders of magnitude as $n_i \to \infty$,

$$\hat{\mathbf{g}}^{\mathrm{a1}}_{bi} = \mathbf{g}_{bi}(\alpha, \beta, \hat{\alpha}_{0i(\alpha,\beta)}) + E^{-1}\left(\frac{\partial h_{2i}}{\partial \alpha_{0i}}\middle|\, \mathbf{d}_{i0}\right) E\left(h_{2i}\frac{\partial \mathbf{g}_{bi}}{\partial \alpha_{0i}}\middle|\, \mathbf{d}_{i0}\right)\Bigg|_{\alpha_{0i}=\hat{\alpha}_{0i(\alpha,\beta)}}$$

$$-\frac{1}{2}E^{-2}\left(\frac{\partial h_{2i}}{\partial \alpha_{0i}}\middle|\, \mathbf{d}_{i0}\right) E\left(\frac{\partial^2 \mathbf{g}_{bi}}{\partial \alpha_{0i}^2}\middle|\, \mathbf{d}_{i0}\right) E(h_{2i}^2|\mathbf{d}_{i0})\Bigg|_{\alpha_{0i}=\hat{\alpha}_{0i(\alpha,\beta)}}, \quad b = 2, 3. \tag{3.1}$$

See Sections E and F of the supplementary material available at *Biostatistics* online for explicit formulae.

Calculation of the expected derivatives appearing in (3.1) requires knowledge of the conditional joint probability of $\mathbf{x}_i$ given $\mathbf{d}_{i0}$, which generally is not available, and of $\mathrm{Var}(\mathbf{d}_{ij}|\mathbf{x}_i, \mathbf{d}_{i0})$ which, as noted in Section 2.2, requires additional modeling assumptions about the intrafamiliar 3-way associations of the responses. By an argument similar to the one used in Wang and Hanfelt (2008), we can safely replace each expectation given on the right-hand side of the equations in Section F of the supplementary material available at *Biostatistics* online by its empirical equivalent, to arrive at a robust adjusted estimating function, say $\hat{\mathbf{g}}^{\mathrm{a}}_{bi}$, $b = 2, 3$, that achieves the same order of bias correction as the less robust adjusted estimating function (3.1). It follows that bias correction can be achieved without requiring any modeling assumptions beyond those needed to write down the original estimating functions (2.3–2.5).

Consider a 2-index asymptotic setting, where both the typical stratum size, q, and the number of the strata, N, may grow to infinity. The unadjusted profile estimating function $\hat{\mathbf{g}}_b$, defined as $\sum_{i=1}^{N} \mathbf{g}_{bi}$ $(\alpha, \beta, \hat{\alpha}_{0i(\alpha,\beta)})$, $b = 2, 3$, has the usual asymptotic distribution as $N \to \infty$, $q \to \infty$, provided that $N/q \to 0$; the corresponding condition for the adjusted estimating function $\hat{\mathbf{g}}^{\mathrm{a}}_b$, defined as $\sum_{i=1}^{N} \hat{\mathbf{g}}^{\mathrm{a}}_{bi}$, $b = 2, 3$, is merely $N^{1/3}/q \to 0$, demonstrating a type of second-order bias correction; see Section G of the supplementary material available at *Biostatistics* online.

Let $\theta = (\alpha^{\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}}$ denote the parameters of interest, and let $\hat{\theta}$ be the root to $\hat{\mathbf{g}}^{\mathrm{a}} = \sum_{i=1}^{N}(\hat{\mathbf{g}}^{\mathrm{aT}}_{2i}, \hat{\mathbf{g}}^{\mathrm{aT}}_{3i})^{\mathrm{T}}$. Let the naive estimator $\hat{\theta}^{\mathrm{p}}$ be the root to a profile estimating function $\hat{\mathbf{s}}^{\mathrm{p}} = (\hat{\mathbf{s}}_2^{\mathrm{T}} + \hat{\mathbf{s}}_1^{\mathrm{T}}, \hat{\mathbf{s}}_3^{\mathrm{T}})^{\mathrm{T}}$, where $\hat{\mathbf{s}}_1 = \sum_{i=1}^{N} \mathbf{s}_{i1}|_{\lambda_i=\hat{\lambda}_{i\alpha}}$ and $\hat{\mathbf{s}}_b = \sum_{i=1}^{N} \mathbf{s}_{ib}|_{\alpha_{0i}=\hat{\alpha}_{0i(\alpha,\beta)}}$, $b = 2, 3$, where $\hat{\lambda}_{i\alpha}$ and $\hat{\alpha}_{0i(\alpha,\beta)}$ are obtained by solving $h_{1i} = 0$ and $h_{2i} = 0$, respectively. By similar arguments to those in Sartori (2003), consistency and asymptotic normality of $\hat{\theta}$ and $\hat{\theta}^{\mathrm{p}}$ can be affected by the relative rates at which the typical stratum size (q) and the number of strata (N) grow. Specifically, the standardized naive estimator $N^{1/2}q^{1/2}(\hat{\theta}^{\mathrm{p}} - \theta)$ converges in distribution to a normal random variable with mean 0 as $N \to \infty, q \to \infty$, provided that $N/q \to 0$; the corresponding condition for the standardized adjusted estimator, namely $N^{1/2}q^{1/2}(\hat{\theta} - \theta)$, is that $N^{1/3}/q \to 0$.

In large samples, the variance of $\hat{\theta}$ can typically be approximated by

$$E\left(\frac{\partial \mathbf{g}}{\partial \theta}\middle|\, \mathbf{d}_0\right)^{-1} \mathrm{Var}(\mathbf{g}|\mathbf{d}_0) E\left(\frac{\partial \mathbf{g}}{\partial \theta}\middle|\, \mathbf{d}_0\right)^{-1^{\mathrm{T}}}\Bigg|_{\theta=\hat{\theta}, \alpha_{0i}=\hat{\alpha}_{0i}}, \tag{3.2}$$

where $\mathbf{d}_0 = \{\mathbf{d}_{10}, \ldots, \mathbf{d}_{N0}\}$ and $\mathbf{g} = \sum_i (\mathbf{g}_{2i}^{\mathrm{T}}, \mathbf{g}_{3i}^{\mathrm{T}})^{\mathrm{T}}$; see Section H of the supplementary material available at *Biostatistics* online for derivation and computation.

3.2 *Combining with the conditional logistic regression score*

The efficiency of the α-estimator obtained from the adjusted versions of $\mathbf{s}_2$ and $\mathbf{s}_3$ may be reduced by not using $\mathbf{s}_1$. We propose to combine the adjusted versions of $\mathbf{s}_2$ and $\mathbf{s}_3$ with the conditional logistic regression score (Breslow *and others*, 1978; Breslow and Day, 1980) based on only the probands' data,

$$\mathbf{s}_{\mathrm{clr}} = \sum_{i=1}^{N} \mathbf{s}_{i\mathrm{clr}} = \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} d_{ij0}\mathbf{x}_{ij0} - \frac{\sum_{k_1,\ldots,k_M \mathrm{in} 1,\ldots,n_i} \left(\sum_{m=1}^{M} \mathbf{x}_{ik_m0}\right) \exp\left\{\left(\sum_{m=1}^{M} \mathbf{x}_{ik_m0}^{\mathrm{T}}\right)\alpha\right\}}{\sum_{k_1,\ldots,k_M \mathrm{in} 1,\ldots,n_i} \exp\left\{\left(\sum_{m=1}^{M} \mathbf{x}_{ik_m0}^{\mathrm{T}}\right)\alpha\right\}} \right],$$

where $M = \sum_{j=1}^{n_i} d_{ij0}$. We obtain a θ-estimator, $\hat{\theta}^{\mathrm{c}}$ say, from the combined estimating function $\hat{\mathbf{g}}^{\mathrm{c}} = (\hat{\mathbf{g}}_2^{\mathrm{aT}} + \mathbf{s}_{\mathrm{clr}}^{\mathrm{T}}, \hat{\mathbf{g}}_3^{\mathrm{aT}})^{\mathrm{T}}$.

It is straightforward to show that the asymptotic properties discussed at the end of Section 3.1 still hold for $\hat{\theta}^{\mathrm{c}}$. In addition, the variance estimator of $\hat{\theta}^{\mathrm{c}}$ still takes the form (3.2) with $\mathbf{g}$ now replaced by $\mathbf{g}^{\mathrm{c}} = (\mathbf{g}_2^{\mathrm{T}} + \mathbf{s}_{\mathrm{clr}}^{\mathrm{T}}, \mathbf{g}_3^{\mathrm{T}})^{\mathrm{T}}$.

4. SIMULATION STUDY

We conducted a simulation study based on the sleep apnea data introduced in Section 1 to evaluate the performance of our method. Note that such a one-to-one matched case–control family study design posed a severe challenge to our proposed estimating function method, and the formal asymptotic theory presented in Section 3 did not strictly apply. Section I of the supplementary material available at *Biostatistics* online gives details of the simulation study.

Let PEF-A and PEF-B denote the unadjusted profile estimating function either including or not including $\mathbf{s}_1$, and let APEF-A and APEF-B refer to the adjusted version either including or not including the conditional score $\mathbf{s}_{\mathrm{clr}}$. We found that APEF-A and APEF-B led to improvements in both the biases of estimators and the coverage rates of 95% confidence intervals. The APEF-A estimator of the regression coefficients in the marginal model (2.1) was more efficient but also more biased than the APEF-B estimator. The APEF-A and APEF-B estimators of the intrafamilial association parameters (2.2) were similar in bias and efficiency. We considered an unadjusted estimating function (PEF-C) naively assuming a common intercept in marginal model (2.1) and found that the PEF-C estimator was the most efficient but had a larger bias than the APEF-A and APEF-B estimators of the parameters in the intrafamilial association model.

5. EXAMPLE: SLEEP APNEA STUDY

We now apply the proposed estimating function to the sleep apnea familial study. The formal asymptotic theory presented in Section 3 does not strictly apply to this matched-pair family study, but the simulation study in Section 4 shows the usefulness of our approach in this type of design.

Our analysis of the familial data from the sleep apnea study was based on 38 case probands with a total of 197 first-degree relatives and 38 matched controls with 160 family members. The family sizes varied from 2 to 13; the total number of relatives in each stratum ranged from 5 to 17. At least one relative was affected with OSA in each stratum. We assumed a logistic regression model for the association between the outcome and covariates, as in (2.1) with a stratum-specific intercept reflecting the matching effect. Covariates considered in the logistic regression model were "age" (years), "gender" (1 for males, 0 for females), and "BMI" (body mass index measured in kg/m^2), where age and BMI were each adjusted to have mean 0. We fitted 2 models for the intrafamilial odds ratio: model I was $r_{ijk0} = \exp(\beta_1)$ and model II was $r_{ijk0} = \exp(\beta_1 + \beta_2 z_{ijk})$, where $z_{ijk} = 0$ if in the ijth family the relationship between the kth member and the proband was parent–child and $z_{ijk} = 1$ if the relationship was sib–sib.

Table 1 shows the results obtained, respectively, from the PEF-A, PEF-B, PEF-C, APEF-A, and APEF-B, where the conditional correlation matrix W_{ij} was set to the identity matrix. The covariates age, BMI, and gender were significantly associated with the presence of sleep apnea at the $\alpha = 0.05$ level.

Table 1. *Results of sleep apnea data analysis*

	PEF-A		PEF-B		PEF-C		APEF-A		APEF-B	
	Estimate (se)	*P*	Estimate (se)	*P*	Estimate (se)	*P*	Estimate (se)	*P*	Estimate (se)	*P*
Using intrafamily log-odds ratio model I										
α_1 : age adjusted	0.087 (0.022)	<0.001	0.076 (0.014)	<0.001	0.066 (0.008)	<0.001	0.071 (0.012)	<0.001	0.070 (0.013)	<0.001
α_2 : BMI adjusted	0.273 (0.059)	<0.001	0.247 (0.042)	<0.001	0.182 (0.019)	<0.001	0.213 (0.033)	<0.001	0.235 (0.038)	<0.001
α_3 : gender	1.503 (0.857)	0.079	0.996 (0.451)	0.027	1.136 (0.198)	<0.001	1.163 (0.389)	0.003	0.942 (0.424)	0.026
Intrafamily log-odds ratio model:										
β_1 : intercept	0.765 (0.501)	0.127	0.814 (0.452)	0.072	0.657 (0.224)	0.003	0.616 (0.439)	0.160	0.601 (0.432)	0.164
Using intrafamily log-odds ratio model II										
α_1 : age adjusted	0.088 (0.023)	<0.001	0.077 (0.015)	<0.001	0.067 (0.008)	<0.001	0.071 (0.012)	<0.001	0.069 (0.012)	<0.001
α_2 : BMI adjusted	0.272 (0.058)	<0.001	0.245 (0.041)	<0.001	0.180 (0.018)	<0.001	0.214 (0.033)	<0.001	0.237 (0.038)	<0.001
α_3 : gender	1.485 (0.847)	0.079	0.973 (0.434)	0.025	1.104 (0.194)	<0.001	1.133 (0.381)	0.003	0.900 (0.409)	0.028
Intrafamily log-odds ratio model:										
β_1 : intercept	0.306 (0.654)	0.640	0.355 (0.580)	0.541	0.139 (0.266)	0.603	0.343 (0.558)	0.539	0.352 (0.532)	0.508
β_2 : relation	1.122 (1.126)	0.319	1.119 (1.044)	0.284	1.159 (0.506)	0.022	0.732 (0.984)	0.457	0.707 (0.971)	0.466

Gender: 1 for male, 0 for female. Relation: 1 for sib–sib, 0 for parent–child. se, standard error.

Older persons, heavier persons, and men were more likely to have sleep apnea than their counterparts. This is consistent with the results from previous analyses by Williamson *and others* (1996) and Wang *and others* (2004). The PEF-A method failed to discover the association of gender with the presence of sleep apnea when controlling for age and BMI. The combined estimating function method (APEF-A) found insufficient evidence to conclude that sleep apnea aggregated within families or a difference between the parent–child association and the sib–sib association. As noted in Section 1, the outcomes of the probands and the relatives were defined differently; the relatives were more likely to be determined as having disease. If we had defined the outcomes of the relatives and the probands in the same way, then the regression coefficients, α, of the marginal mean model would have been smaller than the current estimates.

6. DISCUSSION

This paper proposes a method to analyze finely stratified familial studies with proband sampling that requires modeling only the first 2 joint moments of the data. A simulation study shows the benefits of the approach even in the challenging situation where the probands are individually matched, although the formal asymptotic theory does not apply in this situation.

Our estimator for the nuisance parameter α_{0i} requires at least one affected relative of the probands in each stratum. Therefore, the approach might not be suitable for studies where the number of relatives, $\sum_{j=1}^{n_i} m_{ij}$, is small and the disease is rare among relatives of cases. In such a study, the conditional probability υ_{ijk} might be replaced by

$$\begin{aligned}\upsilon_{ijk}^{\star} &= \mathrm{pr}\left(d_{ijk} = 1 | d_{i10}, \ldots, d_{in_i0}, \sum_{l=1}^{n_i}\sum_{k=1}^{m_{ij}} d_{ilk} > 0\right)\\ &= \mathrm{pr}(d_{ijk} = 1 | d_{ij0}) \Bigg/ \mathrm{pr}\left(\sum_{l=1}^{n_i}\sum_{k=1}^{m_{ij}} d_{ilk} > 0 \,\middle|\, d_{i10}, \ldots, d_{in_i0}\right),\end{aligned}$$

where specification of $\mathrm{pr}\big(\sum_{l=1}^{n_i}\sum_{k=1}^{m_{ij}} d_{ilk} > 0 \big| d_{i10}, \ldots, d_{in_i0}\big)$ would require knowledge of higher-order joint moments of the observations. When this probability is close to 1, $\upsilon_{ijk}^{\star}$ can be approximated by υ_{ijk}. This approximation tends to yield an upward bias in the regression coefficients, α, of the marginal mean model and a bias of uncertain direction in the parameters, β, of the intrafamilial pairwise odds ratio model. In the simulation study summarized in Section 4, we limited the analysis of the relatives' data to the strata with at least one affected relative, and so we implicitly used the approximation $\upsilon_{ijk}^{\star} = \upsilon_{ijk}$. The resulting biases of the estimates from the proposed approach were mild. The probands' data in the strata with no affected relatives contributed to the conditional logistic regression score used in the combined estimating function method (APEF-A).

The general asymptotic properties of the proposed estimator would not change if one were to add an arbitrary nonrandom weight matrix, **D** say, to the combined estimating function, resulting in $\hat{\mathbf{g}}^{\mathrm{c}} = (\hat{\mathbf{g}}_2^{\mathrm{aT}} + \mathbf{D}\mathbf{s}_{\mathrm{clr}}^{\mathrm{T}}, \hat{\mathbf{g}}_3^{\mathrm{aT}})^{\mathrm{T}}$. The choice of D could affect the efficiency of the resulting estimator, however. Determining an optimal weight D in the family data setting considered in this paper, without imposing additional modeling assumptions, is a subject of future research.

Acknowledgments

We thank the editor, associate editor, and a referee for their helpful comments that have improved the manuscript. We also thank Dr Susan Redline at the Rainbow Babies and Children's Hospital of Case Western Reserve University for kindly providing us data from the sleep apnea study supported by a grant from the U.S. National Institutes of Health. *Conflict of Interest:* None declared.

Funding

U.S. National Institutes of Health (R01 MH060628) to J.J.H.

References

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research, Volume 1. The Analysis of Case-Control Studies*. Lyon, France: World Health Organization.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L. and Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology* **108**, 299–307.

Hanfelt, J. J. (2004). Composite conditional likelihood for sparse clustered data. *Journal of the Royal Statistical Society, Series B* **66**, 259–273.

Liang, K. Y. and Beaty, T. H. (2000). Statistical designs for familial aggregation. *Statistical Methods in Medical Research* **9**, 543–562.

Liang, K. Y. and Pulver, A. E. (1996). Analysis of case-control/family sampling design. *Genetic Epidemiology* **13**, 253–270.

Lowe, A. A., Sjoholm, T. T., Ryan, C. F., Fleetham, J. A., Ferguson, K. A. and Remmers, J. E. (2000). Treatment, airway and compliance effects of a titratable oral appliance. *Sleep* **23** (Suppl 4), S172–S178.

Neuhaus, J., Scott, A. J. and Wild, C. J. (2002). The analysis of retrospective family studies. *Biometrika* **89** 23–37.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* **90**, 533–549.

Severini, T. A. (2002). Modified estimating functions. *Biometrika* **89**, 333–343.

Wang, M. and Hanfelt, J. J. (2003). Adjusted profile estimating function. *Biometrika* **90**, 845–858.

Wang, M. and Hanfelt, J. J. (2008). Robust modified profile estimating function with application to the generalized estimating equation. *Journal of Statistical Planning and Inference* **138**, 2029–2044.

Wang, M., Williamson, J. M. and Redline, S. (2004). A semiparametric method for matched case-control family studies with a continuous outcome and proband sampling. *Biometrics* **60**, 644–650.

Whittemore, A. S. (1995). Logistic regression of family data from case-control studies. *Biometrika* **82**, 57–67.

Williamson, J. M., Tosteson, T., Redline, S., Liu, X. and Dawson, D. (1996). Familial aggregation studies with matched proband sampling. *Human Heredity* **46**, 76–84.

Zhao, L. P., Hsu, L., Holte, S., Chen, Y., Quiaoit, F. and Prentice, R. L. (1998). Combined association and aggregation analysis of data from case-control family studies. *Biometrika* **85**, 299–315.

[*Received August 17, 2007; first revision April 29, 2008; second revision August 15, 2008; third revision October 8, 2008; fourth revision October 17, 2008; accepted for publication November 21, 2008*]