Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 931416, 37 pages
doi:10.1155/2012/931416

Research Article

Secondary Analysis under Cohort Sampling Designs Using Conditional Likelihood

Olli Saarela,[1] Sangita Kulathinal,[2,3] and Juha Karvanen[4,5]

[1] *Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada H3A 1A2*
[2] *Indic Society for Education and Development (INSEED), Nashik, Maharashtra 422 011, India*
[3] *Department of Vaccines, National Institute for Health and Welfare, 00271 Helsinki, Finland*
[4] *Department of Mathematics and Statistics, University of Tampere, 33014 Tampere, Finland*
[5] *Department of Mathematics and Statistics, University of Helsinki, 00014 Helsinki, Finland*

Correspondence should be addressed to Olli Saarela, olli.saarela@mcgill.ca

Received 28 July 2011; Revised 29 December 2011; Accepted 24 January 2012

Academic Editor: Kari Auranen

Under cohort sampling designs, additional covariate data are collected on cases of a specific type and a randomly selected subset of noncases, primarily for the purpose of studying associations with a time-to-event response of interest. With such data available, an interest may arise to reuse them for studying associations between the additional covariate data and a secondary non-time-to-event response variable, usually collected for the whole study cohort at the outset of the study. Following earlier literature, we refer to such a situation as secondary analysis. We outline a general conditional likelihood approach for secondary analysis under cohort sampling designs and discuss the specific situations of case-cohort and nested case-control designs. We also review alternative methods based on full likelihood and inverse probability weighting. We compare the alternative methods for secondary analysis in two simulated settings and apply them in a real-data example.

1. Introduction

Cohort sampling designs are two-phase epidemiological study designs where information on time-to-event outcomes of interest over a followup period and some basic covariate data are collected on the whole first-phase study group, referred to as a cohort, and in the second phase, more expensive or difficult-to-obtain additional covariate data are collected only on a subset of the study cohort. This usually comprises the cases, that is, individuals with a disease event of interest during the followup, and a randomly selected subset of noncases. Examples are the case-cohort [1–3] and nested case-control [4, 5] designs. Primarily, such designs are applied for the purpose of studying associations between the time-to-event

outcomes and the covariates collected in the second phase. However, with such data having been collected, an interest frequently arises to reuse it for studying associations between the second-phase covariates and the other available covariate data. For instance, the covariates collected in the second phase could be genotypes, while the other covariates may be various phenotype measurements carried out at the outset of the followup period for the whole cohort. The interest would then be to explain a phenotypic response with the genetic covariates. Following Jiang et al. [6] and Lin and Zeng [7], we refer to such a situation as secondary analysis. Here, we concentrate specifically on non-time-to-event secondary outcomes. Analysis of secondary time-to-event outcomes under the nested case-control design has been considered previously by Saarela et al. [8] and Salim et al. [9].

As our motivating example, we consider here a single cohort which was used in a larger meta-analysis of association between the European lactase persistence genotype and body mass index (BMI) [10], the latter being a secondary outcome in the cohort study in question. The cohort consists of 5073 men aged 55–77 years from southern and western Finland, who originally formed the placebo group of the ATBC cancer prevention study [11]. Whole blood samples of the participants were taken between 1992 and 1993, which is here considered as the baseline of the cohort, with followup for cardiovascular disease events and all-cause mortality available until the end of year 1999. There is no loss to followup, so the only censoring present is of type I due to end of the followup period. This cohort is a part of MORGAM project, an international pooling of cardiovascular cohorts [12]. Genotype data (including the lactase persistence SNP rs4988235) under this project have been collected under a case-cohort design described in detail by [13] and herein in Section 4.3.1. Given such data, our aim is to estimate the association between the lactase persistence genotype and BMI making use of genotype data collected on both the random subcohort and cases of all-cause mortality.

Secondary analysis of case-control data has been studied previously, using profile likelihood [14], inverse selection probability weighting methods [15–17], or retrospective likelihood [6, 7]. However, to the best of our knowledge, a systematic discussion on secondary analysis under cohort sampling designs has been lacking, which we will aim to rectify here by discussing alternative approaches for such an analysis under a generic two-phase study design. We will briefly review the full likelihood approach which utilizes all observed data (Section 2), as well as pseudolikelihoods based on inverse selection probability weighting (Section 3). For these approaches, we propose a conditional likelihood-based alternative (Section 4), restricted to the fully observed second-phase study group. Conditional likelihood inference under cohort sampling designs has been studied previously for the analysis of the primary time-to-event outcome by Langholz and Goldstein [18] and Saarela and Kulathinal [19]; here, we extend these methods to the secondary analysis setting. The main interest is in continuous secondary outcomes, though the approach would also be valid for categorical responses. As special cases of the general setting, we consider case-cohort and nested case-control designs. As extensions to the basic setting, we consider treatment of missing second-phase covariate data and adjustment for left truncation in the case of incident time-to-event outcomes (Section 5). In Section 6, we present two simulation studies, first comparing the efficiencies of the alternative approaches and then demonstrating the potential adverse effects of small sampling fraction in full likelihood inference. We also carry out the analysis in the real-data example using all three alternative methods. As the model for the continuous secondary response variable, in addition to the customary normal distribution, we consider more flexible model specifications, thus aiming to incorporate residual-based model fit diagnostics into the model itself. While other generalizations for the normal distribution have

been proposed, we adopt here the four-parameter normal-polynomial quantile mixture [20], which includes the normal distribution as a special case.

2. Notation, Assumptions, and Full Likelihood

To cover our motivating example and also the general case, we introduce first some notation. Let the set $\mathcal{C} \equiv \{1, \ldots, N\}$ represent the individuals in the cohort. Primary outcome in the study is a time-to-event outcome characterized by random variables (T_i, E_i), $i \in \mathcal{C}$, where T_i corresponds to event time and E_i to the event type of individual i, with $E_i = 0$ indicating a censoring event and $E_i = 1$ a death due to any cause. Extension to incident nonfatal outcomes and multiple outcome types is considered separately in Section 5.2. A secondary outcome of interest Y_i, here BMI, is observed on all study participants at the outset of the study. In addition, there may be other covariates X_i, available on all $i \in \mathcal{C}$, relevant to be included in the analysis. In the present example, these comprise only the age at the start of the followup. Additional covariate data (here the lactase persistence genotype) Z_i are collected only on the second-phase study group $\mathcal{O} \equiv \{i : R_i = 1\} \subseteq \mathcal{C}$, specified by the inclusion indicators $R_i \in \{0, 1\}$, analogously to the survey response/nonresponse setting of Rubin [21]. We will henceforth use vector notations of the type $Z_{\mathcal{O}} \equiv \{Z_i : i \in \mathcal{O}\}$ to represent data obtained on different subsets of individuals. We will not make a distinction between random variables and their realized values in the notation when this is clear from the context. The observed data as a whole are then represented as $(R_{\mathcal{C}}, T_{\mathcal{C}}, E_{\mathcal{C}}, X_{\mathcal{C}}, Y_{\mathcal{C}}, Z_{\mathcal{O}})$. We are interested in the model $P(Y_i \mid X_i, Z_i, \beta)$ for the secondary outcome, more precisely, the parameter describing the association between Y_i and Z_i, which in our case correspond to BMI and the genotypic covariate.

We assume that the first-phase sampling mechanism has been unconfounded in the sense of Rubin [21, page 36], so that we may ignore the first-phase sampling mechanism in all subsequent analyses. This means that the cohort recruitment (possibly through survey sampling) and possible nonresponse depend only on the X_i-covariates; then if all subsequent analyses are conditional on X_i, the selection mechanism may be ignored. In contrast, the second phase sampling may be outcome dependent; the second-phase sampling mechanism is specified by the joint probability distribution for the set of indicator variables $R_{\mathcal{C}}$. This is assumed to be unconfounded with $Z_{\mathcal{C}}$, that is, $R_{\mathcal{C}} \perp (Z_{\mathcal{C}}, \theta) \mid T_{\mathcal{C}}, E_{\mathcal{C}}, X_{\mathcal{C}}, Y_{\mathcal{C}}$. What follows could be further generalized by assuming the sampling mechanism to be ignorable so that $R_{\mathcal{C}} \perp (Z_{\mathcal{C}\setminus\mathcal{O}}, \theta) \mid T_{\mathcal{C}}, E_{\mathcal{C}}, X_{\mathcal{C}}, Y_{\mathcal{C}}, Z_{\mathcal{O}}$, but since most common cohort sampling mechanisms go under the former assumption, and it will simplify the exposition, we will proceed with that. In addition, we assume the random vectors $(T_i, E_i, Y_i, X_i, Z_i)$ either to be (infinitely) exchangeable over unit indices i (cf. [21, page 40]), or equivalently, conditionally independent given the collection of all relevant parameters θ. It should be noted that the exchangeability assumption needs not to be extended to the inclusion indicators R_i. Now, following, for example, Saarela et al. [8], we can write a full (observed data) likelihood expression

$$\begin{aligned} &P(R_{\mathcal{C}}, T_{\mathcal{C}}, E_{\mathcal{C}}, Y_{\mathcal{C}}, Z_{\mathcal{O}} \mid X_{\mathcal{C}}, \theta) \\ &\quad \overset{(\alpha,\beta,\gamma)}{\propto} \prod_{i \in \mathcal{O}} P(T_i, E_i \mid X_i, Y_i, Z_i, \alpha) P(Y_i \mid X_i, Z_i, \beta) P(Z_i \mid X_i, \gamma) \\ &\qquad \times \prod_{i \in \mathcal{C}\setminus\mathcal{O}} \int_{z_i} P(T_i, E_i \mid X_i, Y_i, z_i, \alpha) P(Y_i \mid X_i, z_i, \beta) P(dz_i \mid X_i, \gamma). \end{aligned} \tag{2.1}$$

In the secondary analysis situation, only the parameters β are of interest, while α and γ remain present as nuisance parameters. There are potential drawbacks related to the above likelihood expression; it requires integration over the unobserved covariate data in the set $\mathcal{C} \setminus \mathcal{O}$ (which may be large compared to $\mathcal{O}$), as well as modeling of the population distribution of Z_i. If possible, we would like to avoid this due to the required estimation of the nuisance parameters γ and the risk of model misspecification. Furthermore, observed data likelihoods may become sensitive to misspecification of the model for the response variable; the missing data can act similarly to extra parameters, and the actual model parameters may lose their intended interpretation. This is a real problem especially in cohort sampling designs with a rare event of interest, since the proportion of uncollected covariate data in the study cohort may then be very high. We demonstrate such a situation with simplified simulation example in Section 6.2.

3. Methods Based on Inverse Probability Weighting

Valid estimates for the parameters of the secondary outcome model could alternatively be obtained by using inverses of the first-order inclusion probabilities $P(R_i = 1 \mid T_{\mathcal{C}}, E_{\mathcal{C}}, X_{\mathcal{C}}, Y_{\mathcal{C}})$ (assumed here to be known) as weights (e.g., [22, 23]). The weighted log-likelihood function, approximating the corresponding complete data log-likelihood, to be used for the secondary analysis is

$$\sum_{i \in \mathcal{O}} \frac{\log P(Y_i \mid X_i, Z_i, \beta)}{P(R_i = 1 \mid T_{\mathcal{C}}, E_{\mathcal{C}}, X_{\mathcal{C}}, Y_{\mathcal{C}})} = \sum_{i \in \mathcal{C}} \frac{\mathbf{1}_{\{R_i = 1\}} \log P(Y_i \mid X_i, Z_i, \beta)}{P(R_i = 1 \mid T_{\mathcal{C}}, E_{\mathcal{C}}, X_{\mathcal{C}}, Y_{\mathcal{C}})}. \tag{3.1}$$

In a typical cohort sampling design, all cases would receive unit weights, while noncases selected to the set $\mathcal{O}$ would receive weights greater than one, inverse to their probability to be included in the set of controls/subcohort. While this kind of weighting results in unbiased estimation of the parameters of interest, it will potentially result in reduced efficiency compared to the full and conditional likelihood approaches, since in (3.1) cases receive smaller weights compared to noncases irrespective of whether the case status is actually associated to the secondary outcome or the covariate of interest, and thus the information in these observations may not be fully utilized. We will demonstrate this in a simulation study in Section 6.1. Theoretical justification for estimating function (3.1), as well as variance estimation is discussed in Appendix A. A variation of the above Horvitz-Thompson [24] type of weighting would be poststratification-based estimation (e.g., [15, 25]), with the stratification carried out over the relevant first-phase variables.

4. Conditional Likelihood Inference under Cohort Sampling Designs

4.1. Definition

A very general definition for conditional likelihood is given by Cox and Hinkley [26, pages 16-17] and Cox [27, page 269] as follows. Let U be a random vector corresponding to all observed data, and let this be partitioned into relevant subsets $U = (V, W)$, the transformation not depending on unknown parameters θ. The conditional likelihood given the realization $V = v$ is then the conditional probability distribution $P(W \in dw \mid V = v, \theta)$ with respect to θ.

Few generally accepted rules for choosing the partitioning can be found from the literature, one of these being conditioning on an ancillary statistic or something close to that in order to eliminate nuisance parameters (e.g., [28, 29]). In contrast, although working under the same general definition, here we condition on a sampling mechanism, in order to restrict the analysis into a subgroup for which a useful likelihood expression can be written. Generally such a conditioning may lose information on θ, but will nevertheless produce valid estimates. The rules we set out for choosing the partitioning to construct a conditional likelihood under the two-phase study setting are as follows.

(1) Condition on the sampling mechanism, that is, the set of inclusion indicators R_C, which produced the second-phase study group $\mathcal{O}$ onto which the analysis is to be restricted. Do not condition on any further information on the sampling mechanism, such as inclusion in the subcohort or the set of controls, since this can easily lead to overconditioning which will lose a lot of information. In our notation, such additional information on the sampling mechanism is implicitly included in W and, given the assumptions stated in Section 2, will cancel out of the resulting likelihood expression.

(2) Other observed variables may be placed into V or W at will, depending on the parameters of interest. For instance, if the parameters γ are not of interest, we may condition on $Z_\mathcal{O}$ by placing it in V. If the parameters γ need to be estimated, $Z_\mathcal{O}$ may be placed in W.

(3) We must have $P(W \in dw \mid V = v, \theta)P(V \in dv \mid \theta) = P(U \in du \mid \theta)$, that is, all the relevant observed variables must either be modeled or conditioned upon.

Applying these conditioning rules will reproduce the conditional likelihood expressions obtained previously in special cases of the current framework by Langholz and Goldstein [18] and Saarela and Kulathinal [19]. The proposed conditional likelihood framework also includes the familiar retrospective likelihood, often suggested for analysis under case-control designs (e.g., [30, 31, page 156]), as a special case (see Appendix B). Whereas Langholz and Goldstein [18] considered only the special case of logistic likelihoods, here we aim to first derive the likelihood expressions in the general case before substituting in any specific models.

4.2. Conditional Likelihood Expression: The General Case

Following the stated rules, and making the same general assumptions as in Section 2, we proceed by partitioning the observed data into relevant subsets $W \equiv (T_\mathcal{O}, E_\mathcal{O}, Y_\mathcal{O})$ and $V \equiv (R_C, T_{C\setminus\mathcal{O}}, E_{C\setminus\mathcal{O}}, Y_{C\setminus\mathcal{O}}, X_C, Z_\mathcal{O})$ and, using a shorthand notation of $Q_i \equiv (T_i, E_i, Y_i)$ for all outcome variables, work with a conditional likelihood

$$\begin{aligned} P(Q_\mathcal{O} \mid R_C, Q_{C\setminus\mathcal{O}}, X_C, Z_\mathcal{O}, \theta) &= \frac{P(R_C \mid Q_C, X_C, Z_\mathcal{O}, \theta)P(Q_C, Z_\mathcal{O} \mid X_C, \theta)}{P(R_C \mid Q_{C\setminus\mathcal{O}}, X_C, Z_\mathcal{O}, \theta)P(Q_{C\setminus\mathcal{O}}, Z_\mathcal{O} \mid X_C, \theta)} \\ &\overset{\theta}{\propto} \frac{P(Q_C, Z_\mathcal{O} \mid X_C, \theta)}{P(R_C \mid Q_{C\setminus\mathcal{O}}, X_C, Z_\mathcal{O}, \theta)P(Q_{C\setminus\mathcal{O}}, Z_\mathcal{O} \mid X_C, \theta)}, \end{aligned} \tag{4.1}$$

where the proportionality follows from the assumption on unconfounded second-phase sampling mechanism. In Appendix C, we show that such a conditional likelihood indeed

has the properties of a likelihood, that is, a score function with zero mean and variance equal to the Fisher information. Asymptotic normality of the maximum likelihood estimators based on the conditional likelihood follows obviously from these results in the special case of Bernoulli sampling (Section 4.3.1), although in the general case, this may require further assumptions on the sampling mechanism, such as asymptotic independence of the inclusion indicators R_i.

The ratio of the numerator and the second term in the denominator can be further written as

$$\begin{aligned}\frac{P(Q_{\mathcal{C}}, Z_{\mathcal{O}} \mid X_{\mathcal{C}}, \theta)}{P(Q_{\mathcal{C}\setminus\mathcal{O}}, Z_{\mathcal{O}} \mid X_{\mathcal{C}}, \theta)} &= \frac{\prod_{i\in\mathcal{O}} P(Q_i, Z_i \mid X_i, \theta) \prod_{i\in\mathcal{C}\setminus\mathcal{O}} \int_{z_i} P(Q_i, Z_i \in dz_i \mid X_i, \theta)}{\prod_{i\in\mathcal{O}} \int_{q_i} P(Q_i \in dq_i, Z_i \mid X_i, \theta) \prod_{i\in\mathcal{C}\setminus\mathcal{O}} \int_{z_i} P(Q_i, Z_i \in dz_i \mid X_i, \theta)} \\ &= \prod_{i\in\mathcal{O}} P(Q_i \mid X_i, Z_i, \theta).\end{aligned} \tag{4.2}$$

Here, the product forms follow from the exchangeability assumption for the random vectors (Q_i, X_i, Z_i). Thus, we have obtained

$$P(Q_{\mathcal{O}} \mid R_{\mathcal{C}}, Q_{\mathcal{C}\setminus\mathcal{O}}, X_{\mathcal{C}}, Z_{\mathcal{O}}, \theta) = \frac{\prod_{i\in\mathcal{O}} P(Q_i \mid X_i, Z_i, \theta)}{P(R_{\mathcal{C}} \mid Q_{\mathcal{C}\setminus\mathcal{O}}, X_{\mathcal{C}}, Z_{\mathcal{O}}, \theta)}, \tag{4.3}$$

where the numerator factors into the parametric models $P(Q_i \mid X_i, Z_i, \theta) = P(T_i, E_i \mid X_i, Y_i, Z_i, \alpha) P(Y_i \mid X_i, Z_i, \beta)$. The denominator is the conditional likelihood correction term (sometimes called ascertainment correction, e.g., Ma et al. [32]). Its specific form depends on the second-phase sampling mechanism, and in the general case, it does not reduce into a product form. It depends on the model parameters, since it is not conditioned on all of $(T_{\mathcal{C}}, E_{\mathcal{C}}, Y_{\mathcal{C}}, X_{\mathcal{C}})$. Generally, the challenge in conditional likelihood correction terms is in representing them in terms of parameters estimable from the data. Here, the term can be written as

$$P(R_{\mathcal{C}} \mid Q_{\mathcal{C}\setminus\mathcal{O}}, X_{\mathcal{C}}, Z_{\mathcal{O}}, \theta) = \int_{q_{\mathcal{O}}} P(R_{\mathcal{C}} \mid Q_{\mathcal{C}}, X_{\mathcal{C}}) P(Q_{\mathcal{O}} \in dq_{\mathcal{O}} \mid Q_{\mathcal{C}\setminus\mathcal{O}}, X_{\mathcal{C}}, Z_{\mathcal{O}}, \theta), \tag{4.4}$$

where $P(Q_{\mathcal{O}} \mid Q_{\mathcal{C}\setminus\mathcal{O}}, X_{\mathcal{C}}, Z_{\mathcal{O}}, \theta)$ is given by (4.2), and we have

$$\begin{aligned}P(R_{\mathcal{C}} \mid Q_{\mathcal{C}\setminus\mathcal{O}}, X_{\mathcal{C}}, Z_{\mathcal{O}}, \theta) = \int_{y_i : i\in\mathcal{O}} \int_{t_i : i\in\mathcal{O}} \sum_{e_i : i\in\mathcal{O}} & P(R_{\mathcal{C}} \mid T_{\mathcal{C}}, E_{\mathcal{C}}, Y_{\mathcal{C}}, X_{\mathcal{C}}) \\ & \times \prod_{i\in\mathcal{O}} \left[P(T_i \in dt_i, E_i = e_i \mid X_i, y_i, Z_i, \alpha) P(Y_i \in dy_i \mid X_i, Z_i, \beta) \right].\end{aligned} \tag{4.5}$$

Here, the first term inside the integral specifies the sampling mechanism and is assumed to be known. The obtained likelihood expression utilizes all observed data on covariates Z_i, and it can be easily seen that if there is no association between (T_i, E_i) and Z_i and between Z_i and X_i, we lose no information relevant to learning about the association between Y_i and Z_i, when the

conditional likelihood (4.3) is used instead of the corresponding full likelihood (2.1). This was also demonstrated in the simulation study of Saarela and Kulathinal [19]. On the other hand, if the other observed data do give significant information on the unobserved covariate values $Z_{\mathcal{C}\setminus\mathcal{O}}$, efficiency could potentially be improved by using the full likelihood which utilizes all observed data, with the cost of having to specify a model for Z_i. In Section 6.1, the efficiencies of expressions (2.1) and (4.3) are compared in a simulated setting.

4.3. Special Cases: Case-Cohort and Nested Case-Control Designs

4.3.1. Case-Cohort/Bernoulli Sampling

Here, we are mainly interested in a variation of the "efficient case-cohort design" suggested by Kim and De Gruttola [33, pages 155-156] as an alternative to sampling within strata. To improve efficiency in sampling of the subcohort, here the distribution of some key covariates (in the present notation either $X_{\mathcal{C}}$ or both $(X_{\mathcal{C}}, Y_{\mathcal{C}})$) in the subcohort is approximately matched to that of the cases by first fitting a logistic regression model, say, $\text{logit}\{P(E_i = 1 \mid X_i, \mu)\} = \mu' X_i$, and then selecting the subcohort using Bernoulli sampling with the probabilities $\pi_i \equiv 1/(1 + \exp\{-\widehat{\mu}' X_i\})$, independently of the case status. We then have $P(R_i \mid T_{\mathcal{C}}, E_{\mathcal{C}}, Y_{\mathcal{C}}, X_{\mathcal{C}}, \widehat{\mu}) = P(R_i \mid T_i, E_i, Y_i, X_i, \widehat{\mu})$, where in practice, we make the approximation $P(R_i \mid T_i, E_i, Y_i, X_i, \widehat{\mu}) \approx P(R_i \mid T_i, E_i, Y_i, X_i, \mu)$ and will subsequently suppress the sampling mechanism parameters from the notation. The selection probabilities may also be rescaled to give a desired expected subcohort size m as $\pi_i^* \equiv m\pi_i / \sum_{i\in\mathcal{C}} \pi_i \approx m\pi_i/(NP(E_i = 1))$. More generally, the subcohort selection probability π_i may be any known function of (T_i, E_i, Y_i, X_i), that is, $\pi_i \equiv \pi(T_i, E_i, Y_i, X_i)$, with the sampling design specified by $P(R_{\mathcal{C}} \mid T_{\mathcal{C}}, E_{\mathcal{C}}, Y_{\mathcal{C}}, X_{\mathcal{C}}) = \prod_{i\in\mathcal{C}} P(R_i \mid T_i, E_i, Y_i, X_i) = \prod_{i\in\mathcal{C}} \pi(T_i, E_i, Y_i, X_i)^{R_i} (1 - \pi(T_i, E_i, Y_i, X_i))^{1-R_i}$. The special case of Bernoulli sampling is of interest, because here the product of the first-order inclusion probabilities specifies the joint distribution of the inclusion indicators (see, e.g., [34, pages 62-63]), which considerably simplifies the analysis using conditional likelihood. In the case of stratified without replacement sampling, the following could only be interpreted as a first-order approximation when the sampling fractions are small. In the Bernoulli sampling case, the conditional likelihood correction term reduces into the product form

$$
\begin{aligned}
&P(R_{\mathcal{C}} \mid T_{\mathcal{C}\setminus\mathcal{O}}, E_{\mathcal{C}\setminus\mathcal{O}}, Y_{\mathcal{C}\setminus\mathcal{O}}, X_{\mathcal{C}}, Z_{\mathcal{O}}, \theta) \\
&\quad = \prod_{i\in\mathcal{C}\setminus\mathcal{O}} P(R_i = 0 \mid T_i, E_i, Y_i, X_i) \\
&\qquad \times \int_{y_i : i\in\mathcal{O}} \int_{t_i : i\in\mathcal{O}} \sum_{e_i : i\in\mathcal{O}} \prod_{i\in\mathcal{O}} \big[P(R_i = 1 \mid t_i, e_i, y_i, X_i) \\
&\qquad\qquad \times P(T_i \in dt_i, E_i = e_i \mid X_i, y_i, Z_i, \alpha) P(Y_i \in dy_i \mid X_i, Z_i, \beta)\big] \\
&\quad \overset{(\alpha,\beta)}{\propto} \prod_{i\in\mathcal{O}} \int_{y_i} \int_{t_i} \sum_{e_i} \big[P(R_i = 1 \mid t_i, e_i, y_i, X_i) \\
&\qquad\qquad \times P(T_i \in dt_i, E_i = e_i \mid X_i, y_i, Z_i, \alpha) P(Y_i \in dy_i \mid X_i, Z_i, \beta)\big].
\end{aligned}
\tag{4.6}
$$

Let $\mathcal{E} \equiv \{i : E_i = 1\}$ denote the set of individuals who died during the followup, and let $d \equiv |\mathcal{E}|$. In the case-cohort design discussed in Kulathinal and Arjas [35], Kulathinal et al. [13], and Saarela and Kulathinal [19], all cases are selected to the case-cohort set, so that $P(R_i = 1 \mid T_i, E_i = 1, Y_i, X_i) = 1$, $i \in \mathcal{E}$, while to increase the efficiency of the design, the subcohort is selected with probabilities which depend on the age b_i at the start of the followup (included in the X_i covariates) through a logistic model as discussed above, giving the inclusion probability for non-cases as $P(R_i = 1 \mid T_i, E_i = 0, Y_i, X_i) = \pi(b_i)$, $i \in \mathcal{C} \setminus \mathcal{E}$. Under Bernoulli sampling, the realized sample size $n \equiv |\mathcal{O}|$ is random, with the expected sample size in the present example given by $E(n \mid E_{\mathcal{C}}, X_{\mathcal{C}}) = d + \sum_{i \in \mathcal{C} \setminus \mathcal{E}} \pi(b_i)$. In the case of a mortality outcome and type I censoring at predetermined times c_i, with the observed time given by $T_i \equiv \min(\widetilde{T}_i, c_i)$, where $\widetilde{T}_i$ is the latent time of death, (4.6) further simplifies into (see also [19, pages 12-13])

$$
\begin{aligned}
&P\left(R_{\mathcal{C}} \mid T_{\mathcal{C}\setminus\mathcal{O}}, E_{\mathcal{C}\setminus\mathcal{O}}, Y_{\mathcal{C}\setminus\mathcal{O}}, X_{\mathcal{C}}, Z_{\mathcal{O}}, \theta\right) \\
&\quad \overset{(\alpha,\beta)}{\propto} \prod_{i\in\mathcal{O}} \int_{y_i} \int_{t_i\in[0,c_i]} \left[P\left(R_i = 1 \mid t_i, E_i = 1, y_i, X_i\right) P\left(dt_i, E_i = 1 \mid X_i, y_i, Z_i, \alpha\right)\right. \\
&\qquad\qquad \left. + P\left(R_i = 1 \mid t_i, E_i = 0, y_i, X_i\right) P\left(dt_i, E_i = 0 \mid X_i, y_i, Z_i, \alpha\right)\right] \\
&\qquad \times P\left(dy_i \mid X_i, Z_i, \beta\right) \\
&\quad = \prod_{i\in\mathcal{O}} \int_{y_i} \left[P\left(0 \le T_i < c_i, E_i = 1 \mid X_i, y_i, Z_i, \alpha\right)\right. \\
&\qquad\qquad \left. + \pi(b_i) P\left(T_i \in dc_i, E_i = 0 \mid X_i, y_i, Z_i, \alpha\right)\right] P\left(dy_i \mid X_i, Z_i, \beta\right) \\
&\quad = \prod_{i\in\mathcal{O}} \int_{y_i} \left[P\left(0 \le \widetilde{T}_i < c_i \mid X_i, y_i, Z_i, \alpha\right)\right. \\
&\qquad\qquad \left. + \pi(b_i) P\left(\widetilde{T}_i \ge c_i \mid X_i, y_i, Z_i, \alpha\right)\right] P\left(dy_i \mid X_i, Z_i, \beta\right) \\
&\quad = \prod_{i\in\mathcal{O}} \int_{y_i} \left[1 - (1 - \pi(b_i)) P\left(\widetilde{T}_i \ge c_i \mid X_i, y_i, Z_i, \alpha\right)\right] P\left(dy_i \mid X_i, Z_i, \beta\right).
\end{aligned}
\tag{4.7}
$$

Hence, (4.6) could be represented in terms of the survival function. Suppose now that in addition to the type I censoring due to the end of the followup period, there may be random censoring during the followup; for instance, this may be the case if the outcome in the case-cohort design ($E_i = 1$) is not all-cause mortality, but, say, mortality due to cardiovascular diseases. Deaths due to other causes, denoted as $E_i = 2$, then appear as censoring. As before, let $\widetilde{T}_i$ denote the latent time of death, with E_i indicating the type of death or end of

followup period. Similarly as above, we get

$$\begin{aligned}
&P(R_{\mathcal{C}} \mid T_{\mathcal{C}\setminus\mathcal{O}}, E_{\mathcal{C}\setminus\mathcal{O}}, Y_{\mathcal{C}\setminus\mathcal{O}}, X_{\mathcal{C}}, Z_{\mathcal{O}}, \theta) \\
&\quad = \prod_{i\in\mathcal{O}} \int_{y_i} \big[P\big(0 \le T_i < c_i, E_i = 1 \mid X_i, y_i, Z_i, \alpha\big) \\
&\qquad\qquad + \ \pi(b_i) P\big(0 \le T_i < c_i, E_i = 2 \mid X_i, y_i, Z_i, \alpha\big) \\
&\qquad\qquad + \pi(b_i) P\big(T_i \in dc_i, E_i = 0 \mid X_i, y_i, Z_i, \alpha\big)\big] P\big(dy_i \mid X_i, Z_i, \beta\big) \\
&\quad = \prod_{i\in\mathcal{O}} \int_{y_i} \Big[P\Big(0 \le \widetilde{T}_i < c_i, E_i = 1 \mid X_i, y_i, Z_i, \alpha\Big) \\
&\qquad\qquad + \ \pi(b_i) P\Big(0 \le \widetilde{T}_i < c_i, E_i = 2 \mid X_i, y_i, Z_i, \alpha\Big) \\
&\qquad\qquad + \pi(b_i) P\Big(\widetilde{T}_i \ge c_i \mid X_i, y_i, Z_i, \alpha\Big)\Big] P\big(dy_i \mid X_i, Z_i, \beta\big).
\end{aligned} \tag{4.8}$$

Here, the probabilities $P(0 \le \widetilde{T}_i < c_i, E_i = k \mid X_i, Y_i, Z_i, \alpha)$, $k \in \{1,2\}$, are given by the cumulative incidence functions and $P(\widetilde{T}_i \ge c_i \mid X_i, Y_i, Z_i, \alpha)$ by the joint survival function of a competing risks survival model for the two types of death (e.g., [36, pages 251-252]). These are in principle identifiable from the observed data for the set $\mathcal{O}$, since the subcohort has been selected independently of the case status and thus will include a number of other deaths. On the other hand, if this number is very small, the middle term in the above sum contributes little to the correction term, and thus unstable estimation of the corresponding parameters does not necessarily hinder the estimation of the parameters of interest.

4.3.2. Risk Set Sampling

Consider now a nested case-control sampling mechanism in which all cases are selected to the case-control set with probability one, and m controls per case j are selected using without replacement sampling from the risk set $\mathcal{R}_j \equiv \{i : A_i(T_j) = 1\} \setminus \{j\}$ (of size $r_j \equiv |\mathcal{R}_j| = \sum_{i\in\mathcal{C}} A_i(T_j) - 1$), where $A_i(T_j)$ is the at-risk indicator for subject i at event time T_j. This is carried out independently for all cases $j \in \mathcal{E}$. Let the sampled set of time-matched controls for case j be denoted as $\mathcal{S}_j \subseteq \mathcal{R}_j$, some of which may also be future cases since the sampling is carried out without regard to the future case status of the individuals in the risk set. Let $\mathcal{S}_{\mathcal{E}} \equiv \{\mathcal{S}_j : j \in \mathcal{E}\}$ denote the collection of the sampled risk sets and $\mathcal{S} \equiv \bigcup_{j\in\mathcal{E}} \mathcal{S}_j$ the pooled set of sampled controls. Noting that knowing all of $(T_{\mathcal{C}}, E_{\mathcal{C}})$ specifies the risk sets $\mathcal{R}_{\mathcal{E}} \equiv \{\mathcal{R}_j : j \in \mathcal{E}\}$ as well as the order of the events, the risk set sampling mechanism is specified by the joint probability distribution

$$\begin{aligned}
P(R_{\mathcal{C}} \mid T_{\mathcal{C}}, E_{\mathcal{C}}, Y_{\mathcal{C}}, X_{\mathcal{C}}) &= P(R_{\mathcal{C}} \mid \mathcal{R}_{\mathcal{E}}, \mathcal{E}) \\
&= \sum_{\mathcal{S}_{\mathcal{E}} : \left|\mathcal{S}_j\right| = m, j\in\mathcal{E}} P(R_{\mathcal{C}} \mid \mathcal{R}_{\mathcal{E}}, \mathcal{S}_{\mathcal{E}}, \mathcal{E}) P(\mathcal{S}_{\mathcal{E}} \mid \mathcal{R}_{\mathcal{E}}, \mathcal{E}) \\
&= \sum_{\mathcal{S}_{\mathcal{E}} : \left|\mathcal{S}_j\right| = m, j\in\mathcal{E}} \mathbf{1}_{\{\mathcal{S}\cup\mathcal{E}=\mathcal{O}\}} \prod_{j\in\mathcal{E}} P(\mathcal{S}_j \mid \mathcal{R}_j) \\
&= \prod_{j\in\mathcal{E}} \frac{1}{\binom{r_j}{m}} \sum_{\mathcal{S}_{\mathcal{E}} : \left|\mathcal{S}_j\right| = m, j\in\mathcal{E}} \mathbf{1}_{\{\mathcal{S}\cup\mathcal{E}=\mathcal{O}\}}.
\end{aligned} \tag{4.9}$$

If the sampling has been carried out within strata specified by the covariates Y_C and/or X_C through some function $g(X_i, Y_i)$, the risk sets would be defined within strata as $\mathcal{R}_j = \{i : A_i(T_j) = 1, g(X_i, Y_i) = g(X_j, Y_j)\} \setminus \{j\}$, and the above reasoning would apply similarly with the redefined risk sets. The conditional likelihood correction term now becomes

$$\begin{aligned} &P\left(R_C \mid T_{C\setminus\mathcal{O}}, E_{C\setminus\mathcal{O}}, Y_{C\setminus\mathcal{O}}, X_C, Z_{\mathcal{O}}, \theta\right) \\ &\quad= \int_{y_i: i\in\mathcal{O}} \int_{t_i: i\in\mathcal{O}} \sum_{e_i: i\in\mathcal{O}} \prod_{j\in\mathcal{E}} \frac{1}{\binom{r_j}{m}} \sum_{\mathcal{S}_\mathcal{E}: \left|\mathcal{S}_j\right|=m, j\in\mathcal{E}} \mathbf{1}_{\{\mathcal{S}\cup\mathcal{E}=\mathcal{O}\}} \\ &\qquad\times \prod_{i\in\mathcal{O}} \left[P\left(T_i \in dt_i, E_i = e_i \mid X_i, y_i, Z_i, \alpha\right) P\left(Y_i \in dy_i \mid X_i, Z_i, \beta\right)\right], \end{aligned} \tag{4.10}$$

which does not reduce into a product form. Exact numerical evaluation of this term would require enumeration of all possible combinations of sampled risk sets of size m which would result in the observed case-control set $\mathcal{O}$. This is unlikely to be feasible in practice with realistic sample sizes, which is why consideration of approximations may be necessary. Samuelsen [23] showed the inclusion indicators under risk set sampling to be asymptotically pairwise uncorrelated, with the first-order inclusion probabilities for the noncases $i \in \mathcal{C} \setminus \mathcal{E}$ given by $P(R_i = 1 \mid T_C, E_{C\setminus\{i\}}, E_i = 0) = 1 - \prod_{j\in\mathcal{E}}[1 - mA_i(T_j)/r_j]$. Thus, it might be tempting to replace (4.9) with $\prod_{i\in C} P(R_i \mid T_C, E_C, Y_C, X_C)$. However, the properties of such an approximation will require further research.

5. Extensions

5.1. Missing Second-Phase Covariate Data

Further issue to be considered in practical applications is possible missing covariate data within the set $\mathcal{O}$. For instance, if the covariates to be collected under the cohort sampling design are genotypes, after selection of subjects to be genotyped, it may turn out that the DNA amount or concentration is too low or the genotyping is otherwise unsuccessful. Let $M_\mathcal{O} \equiv \{M_i : i \in \mathcal{O}\}$ be a set of indicator variables indicating whether the measurement of Z_i turned out to be unsuccessful after selection of subject i into the set $\mathcal{O}$, and let $\mathcal{M} \equiv \{i : M_i = 1\} \subseteq \mathcal{O}$. Now Z_i is observed only on the set $\mathcal{O} \setminus \mathcal{M}$. We consider the implications of this separately for all of the three approaches discussed above.

5.1.1. Full Likelihood

If the missingness can be assumed missing at random, that is, $M_\mathcal{O} \perp (Z_\mathcal{M}, \theta) \mid R_C, T_C, E_C, X_C, Y_C, Z_{\mathcal{O}\setminus\mathcal{M}}$, the missingness indicators can again be included in the proportionality constant, and the full likelihood expression becomes

$$\begin{aligned} &P\left(R_C, M_\mathcal{O}, T_C, E_C, Y_C, Z_{\mathcal{O}\setminus\mathcal{M}} \mid X_C, \theta\right) \\ &\quad\overset{(\alpha,\beta,\gamma)}{\propto} \prod_{i\in\mathcal{O}\setminus\mathcal{M}} P(T_i, E_i \mid X_i, Y_i, Z_i, \alpha) P\left(Y_i \mid X_i, Z_i, \beta\right) P\left(Z_i \mid X_i, \gamma\right) \\ &\qquad\times \prod_{i\in C\setminus\mathcal{O}\cup\mathcal{M}} \int_{z_i} P(T_i, E_i \mid X_i, Y_i, z_i, \alpha) P\left(Y_i \mid X_i, z_i, \beta\right) P\left(dz_i \mid X_i, \gamma\right). \end{aligned} \tag{5.1}$$

5.1.2. Inverse Probability Weighting

If the missingness mechanism would be known (and missing at random), a weighted pseudo-likelihood expression for the set $\mathcal{O} \setminus \mathcal{M}$ could be obtained by multiplying the original selection probabilities by the probabilities of observing the value of Z_i after selection

$$\sum_{i \in \mathcal{O} \setminus \mathcal{M}} \frac{\log P(Y_i \mid X_i, Z_i, \beta)}{P(M_i = 0 \mid R_C, T_C, E_C, X_C, Y_C, Z_{\mathcal{O} \setminus \mathcal{M}}) P(R_i = 1 \mid T_C, E_C, X_C, Y_C)}. \tag{5.2}$$

In practice, however, the missingness mechanism would have to be modeled to obtain estimates for these probabilities.

5.1.3. Conditional Likelihood

By partitioning the observed data into $W = (T_{\mathcal{O} \setminus \mathcal{M}}, E_{\mathcal{O} \setminus \mathcal{M}}, Y_{\mathcal{O} \setminus \mathcal{M}})$ and $V = (R_C, M_{\mathcal{O}}, T_{C \setminus \mathcal{O} \cup \mathcal{M}}, E_{C \setminus \mathcal{O} \cup \mathcal{M}}, Y_{C \setminus \mathcal{O} \cup \mathcal{M}}, X_C, Z_{\mathcal{O} \setminus \mathcal{M}})$, we obtain a conditional likelihood

$$\begin{aligned} &P(Q_{\mathcal{O} \setminus \mathcal{M}} \mid R_C, M_{\mathcal{O}}, Q_{C \setminus \mathcal{O} \cup \mathcal{M}}, X_C, Z_{\mathcal{O} \setminus \mathcal{M}}, \theta) \\ &\quad \overset{\theta}{\propto} \frac{P(M_{\mathcal{O}} \mid R_C, Q_C, X_C, Z_{\mathcal{O} \setminus \mathcal{M}}, \theta)}{P(M_{\mathcal{O}} \mid R_C, Q_{C \setminus \mathcal{O} \cup \mathcal{M}}, X_C, Z_{\mathcal{O} \setminus \mathcal{M}}, \theta)} \frac{\prod_{i \in \mathcal{O} \setminus \mathcal{M}} P(Q_i \mid X_i, Z_i, \theta)}{P(R_C \mid Q_{C \setminus \mathcal{O} \cup \mathcal{M}}, X_C, Z_{\mathcal{O} \setminus \mathcal{M}}, \theta)}, \end{aligned} \tag{5.3}$$

where $P(R_C \mid Q_{C \setminus \mathcal{O} \cup \mathcal{M}}, X_C, Z_{\mathcal{O} \setminus \mathcal{M}}, \theta)$ is obtained from (4.5) by replacing the set $\mathcal{O}$ with $\mathcal{O} \setminus \mathcal{M}$. Unlike the second-phase sampling mechanism, the missingness mechanism is generally unknown. From (5.3), it can be seen that if we are willing to assume that $Z_{\mathcal{M}}$ are either missing completely at random or that $M_{\mathcal{O}}$ depends only on X_C, the terms involving the missingness indicators cancel out of the likelihood. On the other hand, if the missingness may depend on the response variables Q_C, the missingness mechanism would have to be modeled. However, in such a case, it may be easier to work with the partitioning $W = (M_{\mathcal{O}}, T_{\mathcal{O}}, E_{\mathcal{O}}, Y_{\mathcal{O}}, Z_{\mathcal{O} \setminus \mathcal{M}})$ and $V = (R_C, T_{C \setminus \mathcal{O}}, E_{C \setminus \mathcal{O}}, Y_{C \setminus \mathcal{O}}, X_C)$ and model the population distribution of Z_i rather than trying to estimate parameters describing the missingness mechanism. This in fact corresponds to what was done by Saarela and Kulathinal [19] and was required there because haplotypes are only partially identifiable from unphased genotype data.

5.2. Incident Outcomes and Left Truncation

Previous discussion was specific to a primary mortality outcome using time on study as the main time scale. In this section, we discuss separately how the different methods can accommodate cohort sampling for incident nonfatal primary outcomes. In the analysis of secondary, non-time-to-event outcomes, the presence of left truncation due to exclusion of cases of prevalent disease presents an additional complication. If the parameters of the secondary outcome model correspond to the background population alive at the cohort baseline (rather than to the disease-free population), this additional selection factor requires further adjustment. If the primary outcome is a mortality endpoint, this is not an issue, since then there is no further selection due to prevalent conditions. In likelihood-based adjustment for left truncation, the main time scale of the analysis has to be chosen as age instead of time on

study. In survival modeling, it is well known that conditioning on event-free survival until the age at study baseline corresponds to exclusion of the followup time before that (e.g., [37, page 580], and the references therein). However, this is no longer true in the case of missing covariate data (e.g., [8, pages 5997-5998]), or indeed in the analysis of secondary outcomes, as will be demonstrated below. The set notations $\mathcal{C}$ and $\mathcal{O}$ are here taken to refer to the disease free cohort and second-phase study group.

It should also be noted that under case-cohort designs it is common to collect second-phase covariate data for more than a single outcome, since the case-cohort design naturally enables the analysis of multiple outcomes using a single subcohort selection. This is also the case in our example cohort discussed in Sections 1 and 6.3, where, in addition to cases of all-cause mortality, genotype data has been collected also on cases of nonfatal incident cardiovascular disease events. To keep the example simple, in the data analysis, we consider only the case-cohort set for all-cause mortality. However, it is in principle straightforward to accommodate multiple outcome types in the likelihood expressions discussed below by using competing risks type notation where T_i denotes the observed time of the first incident disease event, death, or censoring, with $E_i = 0$ indicating censoring and $E_i \in \{1, 2, \ldots, K\}$ the different types of outcome events for which the second-phase covariate data is collected. Since utilizing the likelihood expressions does not necessitate estimation of K different hazard functions, the endpoint definitions may be pooled as seen suitable.

5.2.1. Full Likelihood

The full likelihood expression (2.1) is now conditioned on the selection rule $T_{\mathcal{C}} \geq b_{\mathcal{C}} \equiv T_i \geq b_i$ for all $i \in \mathcal{C}$, that is, event-free survival until the age at the cohort baseline. The likelihood expression becomes

$$\begin{aligned} &P(R_{\mathcal{C}}, T_{\mathcal{C}}, E_{\mathcal{C}}, Y_{\mathcal{C}}, Z_{\mathcal{O}} \mid T_{\mathcal{C}} \geq b_{\mathcal{C}}, X_{\mathcal{C}}, \theta) \\ &\overset{(\alpha,\beta,\gamma)}{\propto} \prod_{i\in\mathcal{O}} \frac{P(T_i, E_i \mid X_i, Y_i, Z_i, \alpha) P(Y_i \mid X_i, Z_i, \beta) P(Z_i \mid X_i, \gamma)}{\int_{y_i} \int_{z_i} P(T_i \geq b_i \mid X_i, y_i, z_i, \alpha) P(\mathrm{d}y_i \mid X_i, z_i, \beta) P(dz_i \mid X_i, \gamma)} \\ &\quad \times \prod_{i\in\mathcal{C}\setminus\mathcal{O}} \frac{\int_{z_i} P(T_i, E_i \mid X_i, Y_i, z_i, \alpha) P(Y_i \mid X_i, z_i, \beta) P(dz_i \mid X_i, \gamma)}{\int_{y_i} \int_{z_i} P(T_i \geq b_i \mid X_i, y_i, z_i, \alpha) P(dy_i \mid X_i, z_i, \beta) P(dz_i \mid X_i, \gamma)}. \end{aligned} \tag{5.4}$$

5.2.2. Inverse Probability Weighting

The weighted pseudolikelihood approach does not readily take into account additional selection which occurs in studies of incident outcomes. This was also pointed out by Reilly et al. [15], who in their discussion note that their reweighting approach for case-control studies is not valid for incident or nested case-control studies, since the incident cases and healthy controls available cannot be reweighted to represent the general population. Although this is true for the basic estimating function (3.1), similarly as in the missing data case, a double weighting mechanism can be devised to account for the additional selection step as

$$\sum_{i\in\mathcal{O}} \frac{\log P(Y_i \mid X_i, Z_i, \beta)}{P(T_i \geq b_i \mid X_i, Y_i, Z_i, \widehat{\alpha}) P(R_i = 1 \mid T_{\mathcal{C}}, E_{\mathcal{C}}, X_{\mathcal{C}}, Y_{\mathcal{C}})}. \tag{5.5}$$

Since this weighting requires estimation of the time-to-event model parameters α, resampling-based variance estimation would be required to obtain valid standard errors for the estimates of β to account for the uncertainty in estimation of the weights.

5.2.3. Conditional Likelihood

As in Section 4.2, we get

$$P\left(Q_{\mathcal{O}} \mid R_C, T_C \geq b_C, Q_{C\setminus\mathcal{O}}, X_C, Z_{\mathcal{O}}, \theta\right) = \frac{\prod_{i\in\mathcal{O}} P(Q_i \mid X_i, Z_i, \theta)}{P\left(R_C, T_C \geq b_C \mid Q_{C\setminus\mathcal{O}}, X_C, Z_{\mathcal{O}}, \theta\right)}, \tag{5.6}$$

where the correction term becomes

$$\begin{aligned} &P\left(R_C, T_C \geq b_C \mid Q_{C\setminus\mathcal{O}}, X_C, Z_{\mathcal{O}}, \theta\right) \\ &\quad = \int_{y_i: i\in\mathcal{O}} \int_{t_i\in[b_i,\infty): i\in\mathcal{O}} \sum_{e_i: i\in\mathcal{O}} P(R_C \mid T_C, E_C, Y_C, X_C) \\ &\qquad \times \prod_{i\in\mathcal{O}} \left[\left(T_i \in dt_i, E_i = e_i \mid X_i, y_i, Z_i, \alpha\right) P\left(Y_i \in dy_i \mid X_i, Z_i, \beta\right)\right]. \end{aligned} \tag{5.7}$$

In the case-cohort design discussed in Section 4.3.1, with type I censoring, this further simplifies into

$$\begin{aligned} &P\left(R_C, T_C \geq b_C \mid T_{C\setminus\mathcal{O}}, E_{C\setminus\mathcal{O}}, Y_{C\setminus\mathcal{O}}, X_C, Z_{\mathcal{O}}, \theta\right) \\ &\quad \overset{(\alpha,\beta)}{\propto} \prod_{i\in\mathcal{O}} \int_{y_i} \int_{t_i\in[b_i,c_i]} \sum_{e_i} P\left(R_i = 1 \mid t_i, e_i, y_i, X_i\right) \\ &\qquad \times P\left(T_i \in dt_i, E_i = e_i \mid X_i, y_i, Z_i, \alpha\right) P\left(Y_i \in dy_i \mid X_i, Z_i, \beta\right) \\ &\quad = \prod_{i\in\mathcal{O}} \int_{y_i} \left[P\left(b_i \leq \tilde{T}_i < c_i \mid X_i, y_i, Z_i, \alpha\right) + \pi(b_i) P\left(\tilde{T}_i \geq c_i \mid X_i, y_i, Z_i, \alpha\right)\right] \\ &\qquad \times P\left(Y_i \in dy_i \mid X_i, Z_i, \beta\right) \\ &\quad = \prod_{i\in\mathcal{O}} \int_{y_i} \left[P\left(\tilde{T}_i \geq b_i \mid X_i, y_i, Z_i, \alpha\right) - (1 - \pi(b_i)) P\left(\tilde{T}_i \geq c_i \mid X_i, y_i, Z_i, \alpha\right)\right] \\ &\qquad \times P\left(Y_i \in dy_i \mid X_i, Z_i, \beta\right). \end{aligned} \tag{5.8}$$

6. Illustrations

6.1. Simulation Study

In order to compare the efficiency of the alternative estimation methods discussed above, namely, full likelihood, conditional likelihood, and weighted pseudolikelihood, we supplemented the cohort data described in Section 1 with a simulated covariate, following

the real-data-based simulation approach described by Saarela et al. in [8, pages 5998–6000], in order to have the simulation setting resemble as closely as possible the real data analysis setting of Section 6.3. This Bayesian procedure corresponds to multiple imputation [21], where given the observed data (T_C, E_C, X_C, Y_C) and predetermined associations with the primary and secondary outcomes, an additional "missing" covariate Z_C is simulated from the posterior predictive distribution $P(Z_C \mid T_C, E_C, X_C, Y_C)$. Case-cohort sampling was then simulated in the complete datasets $(T_C, E_C, X_C, Y_C, Z_C)$ thus obtained. The objective of this approach, as compared to using completely simulated data, was to obtain simulation results directly relevant to the real study of interest. We included a set of $N = 5039$ individuals with a BMI measurement (Y_i) available and eligible for the case-cohort study. The primary (time-to-event) outcome (T_i, E_i) was taken to be all-cause mortality, with $d = 996$. Given the primary and secondary outcome data and age at baseline $(X_i \equiv b_i)$ observed for the cohort, and the current parameter values, additional binary covariate values were drawn from the conditional distributions

$$P(Z_i = z_i \mid T_i, E_i, Y_i, X_i, \alpha, \beta, \gamma) = \frac{P(T_i, E_i \mid X_i, Y_i, z_i, \alpha) P(Y_i \mid X_i, z_i, \beta) P(Z_i = z_i \mid X_i, \gamma)}{\sum_{z_i \in \{0,1\}} P(T_i, E_i \mid X_i, Y_i, z_i, \alpha) P(Y_i \mid X_i, z_i, \beta) P(Z_i = z_i \mid X_i, \gamma)}, \tag{6.1}$$

where the model for the survival outcome, $P(T_i \in dt_i, E_i = e_i \mid X_i, Y_i, Z_i, \alpha) \overset{\alpha}{\propto} [\lambda_i(t_i)]^{\mathbf{1}_{\{e_i=1\}}} S_i(t_i)$, was specified as a proportional hazards model $\lambda_i(u) \equiv \lambda_0(u) \exp\{\alpha_1 b_i + \alpha_2 Y_i + \alpha_3 Z_i\}$, with time on study as the time scale and age at baseline as a covariate. Here, the the baseline hazard function λ_0 was specified in terms of a piecewise constant function using 15 time bins of equal length over the followup period of seven years. Flat improper priors were used for all parameters (those not fixed to selected values), on log scale for the nonnegative parameters.

Normal model $Y_i \mid X_i, Z_i, \beta \sim N(\beta_0 + \beta_1 b_i + \beta_2 Z_i, \beta_3^2)$ was used for the secondary outcome, while the population distribution of the additional covariate was specified as $P(Z_i = z_i \mid X_i, \gamma) = \gamma^{z_i}(1-\gamma)^{1-z_i}$. Here, parameters α_3, β_2, and γ were fixed at selected values while the other parameters were allowed to be determined by the data and integrated out by drawing from their respective full conditional posterior distributions using Markov chain Monte Carlo sampling. Variables Y_i and b_i were centered but otherwise untransformed. 1000 values for each Z_i, $i \in C$, were obtained by running the MCMC sampler for 25000 rounds after a 10000-round burn-in with each set of fixed values of $(\alpha_3, \beta_2, \gamma)$ given in Table 1 and saving every 25th state of the chain. In each of the complete datasets obtained by combining the observed data and the simulated covariate values Z_i, case-cohort selection was carried out by matching the age distribution of the subcohort to that of the cases by first fitting a logistic regression model with age as a covariate to the observed survival status, as detailed in Section 4.3.1. The predictive probabilities from the logistic model were scaled to give an expected subcohort size of 1000, and the subcohort was selected using the obtained probabilities in Bernoulli sampling. This procedure gave an expected case-cohort set size of $E(n \mid E_C, X_C) = 1761$.

In fitting models to the datasets so obtained, we used the same model specifications as above, with the exception of the proportional hazards model, where we fitted a "misspecified" Weibull model $\lambda_i(u) \equiv (\alpha_4/\alpha_0)(u/\alpha_0)^{\alpha_4 - 1} \exp\{\alpha_1 b_i + \alpha_2 Y_i + \alpha_3 Z_i\}$. Even though this is a different model than the one specified for the simulation step, we did not simulate the time-to-event outcome data, which was taken from the example cohort; a comparison between

Table 1: Maximum likelihood estimates of the parameters α of the all-cause mortality model $\lambda_i(t) = (\alpha_4/\alpha_0)(t/\alpha_0)^{\alpha_4-1}\exp\{\alpha_1 b_i + \alpha_2 Y_i + \alpha_3 Z_i\}$, β of the BMI model $Y_i \mid X_i, Z_i, \beta \sim N(\beta_0 + \beta_1 b_i + \beta_2 Z_i, \beta_3^2)$, and γ, the population frequency of the simulated covariate (true value 0.25). Parameters of main interest are α_2 and β_2, which characterize the association between the simulated covariate and primary (all-cause mortality) and secondary (BMI) outcomes, respectively. The values are means and standard deviations over 1000 replications. SE stands for the standard errors of the maximum likelihood estimates.

α_3	β_2	Lik.	$\widehat{\alpha}_0$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_3$	SD($\widehat{\alpha}_3$)	SE($\widehat{\alpha}_3$)	$\widehat{\alpha}_4$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	SD($\widehat{\beta}_2$)	SE($\widehat{\beta}_2$)	$\widehat{\beta}_3$	$\widehat{\gamma}$
0.0	0.0	Full	23.72	0.09	−0.01	−0.01	0.10	0.10	1.33	0.00	−0.08	0.01	0.22	0.22	4.10	0.25
		Cond.	23.74	0.09	−0.01	0.00	0.10	0.10	1.33	0.00	−0.08	0.00	0.23	0.23	4.17	—
		Weighted	—	—	—	—	—	—	—	−0.01	−0.08	0.03	0.30	0.30	4.10	—
0.5	0.0	Full	26.30	0.09	−0.01	0.49	0.09	0.09	1.33	0.00	−0.08	0.01	0.22	0.21	4.10	0.25
		Cond.	26.33	0.09	−0.01	0.50	0.09	0.09	1.33	0.00	−0.08	0.00	0.22	0.22	4.17	—
		Weighted	—	—	—	—	—	—	—	−0.01	−0.08	0.03	0.30	0.30	4.10	—
0.0	0.5	Full	23.73	0.09	−0.01	−0.01	0.10	0.10	1.33	−0.13	−0.08	0.53	0.25	0.23	4.09	0.25
		Cond.	23.75	0.09	−0.01	0.00	0.10	0.10	1.33	−0.13	−0.08	0.52	0.24	0.23	4.17	—
		Weighted	—	—	—	—	—	—	—	−0.13	−0.08	0.53	0.31	0.31	4.09	—
0.5	0.5	Full	26.32	0.09	−0.02	0.49	0.09	0.09	1.33	−0.14	−0.08	0.55	0.23	0.22	4.09	0.25
		Cond.	26.34	0.09	−0.01	0.50	0.09	0.09	1.33	−0.14	−0.08	0.54	0.22	0.22	4.17	—
		Weighted	—	—	—	—	—	—	—	−0.14	−0.08	0.54	0.31	0.31	4.09	—
1.0	0.0	Full	29.62	0.09	−0.01	0.99	0.09	0.09	1.35	0.00	−0.08	0.02	0.21	0.21	4.10	0.25
		Cond.	29.64	0.09	−0.01	1.00	0.09	0.09	1.35	0.00	−0.08	0.01	0.22	0.21	4.17	—
		Weighted	—	—	—	—	—	—	—	−0.01	−0.08	0.03	0.29	0.29	4.10	—
0.0	1.0	Full	23.77	0.09	−0.01	0.00	0.10	0.10	1.33	−0.28	−0.08	1.11	0.26	0.23	4.07	0.25
		Cond.	23.77	0.09	−0.01	0.00	0.10	0.10	1.33	−0.27	−0.08	1.06	0.24	0.23	4.15	—
		Weighted	—	—	—	—	—	—	—	−0.26	−0.08	1.05	0.32	0.32	4.07	—
1.0	1.0	Full	29.67	0.09	−0.03	1.00	0.09	0.09	1.35	−0.29	−0.08	1.14	0.24	0.22	4.07	0.25
		Cond.	29.65	0.09	−0.02	1.00	0.09	0.09	1.35	−0.28	−0.08	1.08	0.22	0.21	4.14	—
		Weighted	—	—	—	—	—	—	—	−0.27	−0.08	1.07	0.31	0.31	4.07	—

fitted piecewise constant and Weibull hazards (results not shown) indicated that the Weibull model is adequate for modeling these data, accounting for the increasing baseline mortality rate due to ageing of the cohort. Full likelihood function (2.1), conditional likelihood function (4.3), and weighted pseudolikelihood function (3.1) were maximized with respect to the parameters (α, β, γ), (α, β), and β, respectively, substituting in the above parametric models. The numerical optimization was carried out using the `optim` function of the *R* statistical software, applying the BFGS optimization algorithm [38]. The integrals over y_i in (4.7) were evaluated using numerical (quadrature) integration, using the `gsl_integration_qagi` function of GNU scientific library [39]. The standard errors of the maximum likelihood estimators were evaluated by inverting the numerically differentiated Hessian matrix at the maximum likelihood point. The standard errors of the inverse-probability-weighted estimates were evaluated using the robust variance formula of Appendix A with the observed information (A.4) and observed score (A.5). The results from the 1000 replications are presented in Table 1. Here, the parameters $(\alpha_3, \beta_2, \gamma)$, corresponding to the simulated covariate, are of main interest; the other parameters reflect the observed data, and thus the Monte Carlo variances for these are not relevant. The standard errors for the corresponding parameters are given in the real-data example of Section 6.3. The results indicate that conditional likelihood estimation gives similar efficiency compared to the corresponding full likelihood estimates irrespective of whether the covariate of interest was associated with the survival outcome. In contrast, both likelihood-based approaches gave better efficiency compared to inverse probability weighting. This is expected, as the weighted pseudolikelihood gives smaller weights to cases, whereas the conditional likelihood weights cases and noncases differentially only to the extent they actually differ. Both the inverse observed information and robust variance estimates agreed well with the Monte Carlo variances.

6.2. Multimodality under Full Likelihood When the Sampling Fraction Is Small

Let now the observed data be only (R_C, Y_C, Z_O), and let the sampling mechanism be simple random sampling without replacement; $P(R_C) = 1/\binom{N}{n}$, where the sample size n is fixed. The specific aim of the following example is to demonstrate the vulnerability of observed data likelihoods, integrated over the missing covariate data on the set $C \setminus O$, to misspecification of the response model. The full likelihood expression now becomes $P(R_C, Y_C, Z_O \mid \theta) \overset{(\beta,\gamma)}{\propto} \prod_{i\in O}[P(Y_i \mid Z_i, \beta)P(Z_i \mid \gamma)]\prod_{i\in C\setminus O}\int_{z_i} P(Y_i \mid z_i, \beta)P(Z_i \in dz_i \mid \gamma)$. Due to the simple random sampling, the corresponding conditional likelihoods simplify into $P(Y_O \mid R_C, Y_{C\setminus O}, Z_O, \theta) = \prod_{i\in O} P(Y_i \mid Z_i, \beta)$ and $P(Y_O, Z_O \mid R_C, Y_{C\setminus O}, \theta) = \prod_{i\in O}[P(Y_i \mid Z_i, \beta)P(Z_i \mid \gamma)]$.

With $N = 1000$, we simulated covariate values from $Z_i \sim$ Bernoulli(0.2) and two different sets of response values (both independently of Z_i) from $Y_i \sim N(4, 4)$ and $Y_i \sim$ Gamma$(4, 1)$ (same mean and variance, but the latter distribution is skewed to the right). We fit two alternative models to these data, in both models $Z_i \mid \gamma \sim$ Bernoulli(γ), with the models for the response specified as $Y_i \mid Z_i, \beta \sim N(\beta_0 + \beta_1 Z_i, \beta_2^2)$ and $Y_i \mid Z_i, \beta \sim$ NPQM$(\beta_0 + \beta_1 Z_i, \beta_2, \beta_3, \beta_4)$, where β are the collections of all model parameters. The latter model is the normal-polynomial quantile mixture distribution proposed by Karvanen ([20, pages 950–953]; see also *R* package `Lmoments`), which we apply here for regression modeling. The parametrization here can be expressed in terms of the first four L-moments as $\lambda_1 = E(Y_i \mid Z_i, \beta) = \beta_0 + \beta_1 Z_i$, $\lambda_2 = \beta_2/\sqrt{\pi}$ (L-scale), $\lambda_3 = \beta_3$ (L-skewness), and $\lambda_4 = \beta_4$ (L-kurtosis). This distribution is suitable for regression modeling as the first L-moment is the mean of

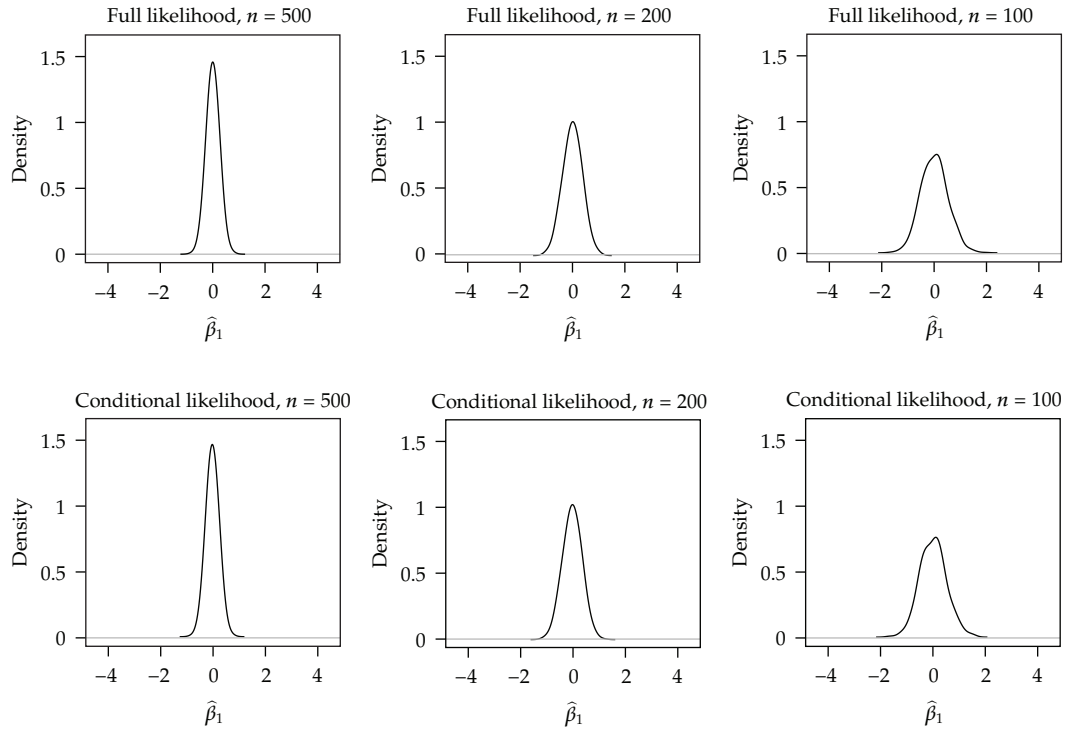

Figure 1: Sampling distributions for maximum likelihood estimates of regression coefficient β_1 from 1000 replications; normal model fitted to normal data ($Y_i \sim N(4, 4)$).

the distribution, while the two additional shape parameters allow for more flexible modeling of the residual distribution. The NPQM distribution includes normal distribution as a special case when $\lambda_3 = 0$ and $\lambda_4/\lambda_2 = 0.1226$. The full and conditional log-likelihood expressions under these model specifications were maximized with respect to β and γ as described in the previous section. Initial values for the algorithm were set at the sample moments calculated from the marginal distribution of the response variable, with the regression coefficient set to $\beta_1 = 0$ and $\gamma = 0.5$.

Table 2 shows the mean maximum likelihood estimates over 1000 replications for all 8 combinations of data generating model, fitted model, and estimation method. With the response simulated from normal distribution, both models estimated using either full or conditional likelihood give the expected results (Figures 1 and 2). However, with the response simulated from gamma distribution and the sampling fraction n/N small, the misspecified normal model estimated using full likelihood indicates a spurious association between the response and the covariate. This is because the missing data act as extra parameters, allowing the optimization procedure to obtain a better fit to the skewed data. The corresponding sampling distributions for β_1 shown in Figure 3 have become bimodal. It should be noted that the initial values given to the optimization algorithm corresponded always to the "correct solution" of no covariate effect, thus enabling the algorithm to find the correct mode. However, bimodality in the sampling distribution does not necessarily indicate that the likelihood functions given a single realized dataset would be bimodal.

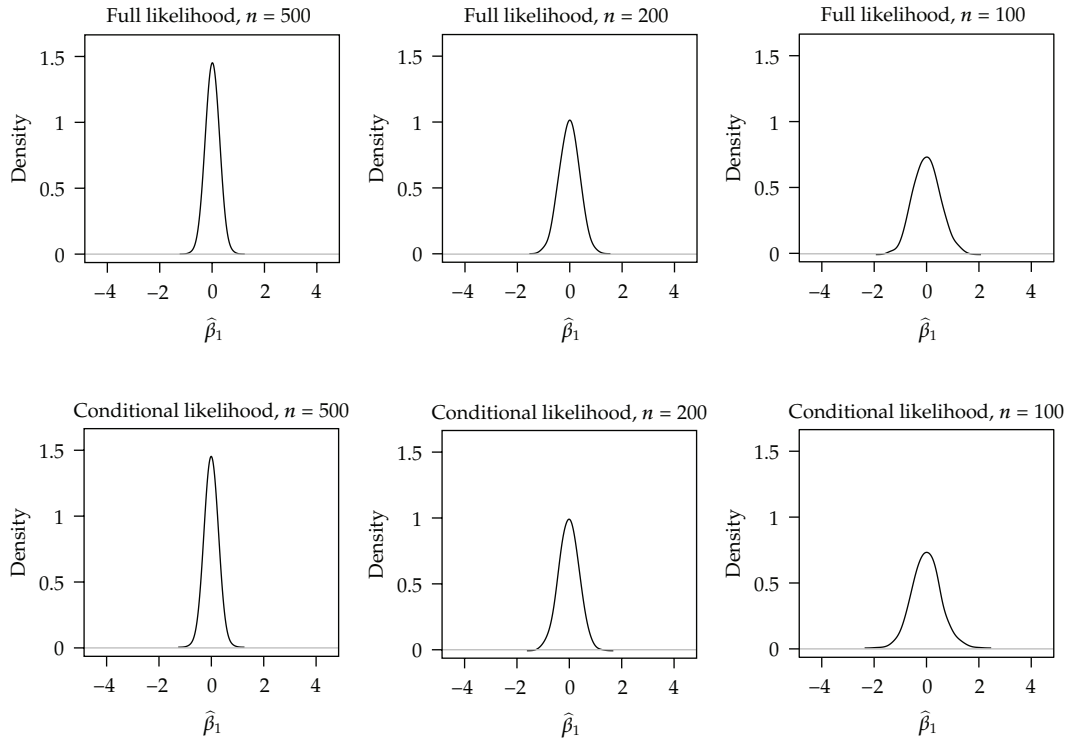

Figure 2: Sampling distributions for maximum likelihood estimates of regression coefficient β_1 from 1000 replications; NPQM model fitted to normal data ($Y_i \sim N(4,4)$).

Using the NPQM model which allows a skewed residual distribution does not correct the situation either when combined with full likelihood estimation (Figure 4). This is not unexpected since adding more parameters to an already overparameterized situation is not necessarily helpful in solving identification problems. In contrast, conditional likelihood continues to give reasonable results in all cases, with the skewness of the data correctly reflected by the parameter β_3 of the NPQM model. Some degree of misspecification of the response model is required for the multimodality to appear since when we simulated the response variable from NPQM distribution, the full likelihood fit of the NPQM model indicated no problems, while the full likelihood fit of the normal model showed the same problems as with gamma data (Figures 5 and 6). We also repeated all the simulation alternatives with $\beta_1 = 1$ ($Y_i \sim Z_i + N(4,4)$ and $Y_i \sim Z_i + \text{Gamma}(4,1)$), with the conclusions essentially unchanged (Table 3 and Figures 7, 8, 9, and 10).

6.3. An Example with Real Data

The case-cohort set for all-cause mortality in the example cohort ($N = 5039$) is of size $n = 1816$, the union of a subcohort of size 1068 and 996 deaths due to any cause (since the subcohort has been selected independently of the case status, it includes a number of cases). The case-cohort design applied in this cohort was described in Section 4.3.1. Due to

Table 2: Maximum likelihood estimates of parameters β of the outcome models $Y_i \mid Z_i, \beta \sim N(\beta_0 + \beta_1 Z_i, \beta_2^2)$ and $Y_i \mid Z_i, \beta \sim \text{NPQM}(\beta_0 + \beta_1 Z_i, \beta_2, \beta_3, \beta_4)$ and the population frequency γ of the binary covariate Z_i (means over 1000 replications). Outcome variable was simulated from $Y_i \sim N(4, 4)$ and $Y_i \sim \text{Gamma}(4, 1)$.

Normal data	n	Normal model					NPQM model							
		$\widehat{\beta}_0$	$\widehat{\beta}_1$	SE($\widehat{\beta}_1$)	$\widehat{\beta}_2$	$\widehat{\gamma}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	SE($\widehat{\beta}_1$)	$\widehat{\beta}_2$	$\widehat{\beta}_3$	SE($\widehat{\beta}_3$)	$\widehat{\beta}_4$	$\widehat{\gamma}$
Full lik.	1000	4.00	0.00	0.16	2.00	0.20	4.00	0.00	0.16	2.00	0.00	0.01	0.14	0.20
	500	4.00	0.01	0.22	1.99	0.20	4.00	0.01	0.22	1.99	0.00	0.01	0.14	0.20
	200	4.00	0.00	0.35	1.99	0.20	4.00	0.00	0.35	1.99	0.00	0.01	0.14	0.20
	100	4.00	0.01	0.50	1.99	0.20	4.00	0.01	0.50	1.99	0.00	0.02	0.14	0.20
Cond. lik.	500	4.00	0.01	0.22	1.99	0.20	4.00	0.01	0.22	1.99	0.00	0.02	0.14	0.20
	200	4.01	0.00	0.35	1.99	0.20	4.01	0.00	0.35	1.99	0.00	0.03	0.13	0.20
	100	4.00	0.01	0.50	1.98	0.20	4.00	0.01	0.50	1.98	0.00	0.05	0.13	0.20
Gamma data	n	Normal model					NPQM model							
		$\widehat{\beta}_0$	$\widehat{\beta}_1$	SE($\widehat{\beta}_1$)	$\widehat{\beta}_2$	$\widehat{\gamma}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	SE($\widehat{\beta}_1$)	$\widehat{\beta}_2$	$\widehat{\beta}_3$	SE($\widehat{\beta}_3$)	$\widehat{\beta}_4$	$\widehat{\gamma}$
Full lik.	1000	4.00	0.01	0.16	1.99	0.20	3.99	0.00	0.13	2.01	0.15	0.01	0.14	0.20
	500	4.00	0.01	0.22	1.99	0.20	3.99	0.01	0.20	2.01	0.15	0.01	0.14	0.20
	200	3.97	0.15	0.38	1.98	0.20	3.98	0.07	0.35	2.00	0.15	0.02	0.14	0.20
	100	3.86	0.91	0.38	1.88	0.19	3.95	0.18	0.32	1.96	0.16	0.02	0.14	0.22
Cond. lik.	500	4.00	0.00	0.22	1.99	0.20	4.00	0.00	0.19	2.01	0.15	0.02	0.14	0.20
	200	4.00	−0.01	0.35	1.98	0.20	4.01	0.00	0.29	2.00	0.15	0.03	0.14	0.20
	100	4.00	−0.01	0.50	1.97	0.20	4.01	0.01	0.41	2.00	0.15	0.05	0.13	0.20

Table 3: Maximum likelihood estimates of parameters β of the outcome models $Y_i \mid Z_i, \beta \sim N(\beta_0 + \beta_1 Z_i, \beta_2^2)$ and $Y_i \mid Z_i, \beta \sim \text{NPQM}(\beta_0 + \beta_1 Z_i, \beta_2, \beta_3, \beta_4)$ and the population frequency γ of the binary covariate Z_i (means over 1000 replications). Outcome variable was simulated from $Y_i \mid Z_i \sim Z_i + N(4, 4)$ and $Y_i \mid Z_i \sim Z_i + \text{Gamma}(4, 1)$.

Normal data	n	Normal model					NPQM model							
		$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\text{SE}(\widehat{\beta}_1)$	$\widehat{\beta}_2$	$\widehat{\gamma}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\text{SE}(\widehat{\beta}_1)$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\text{SE}(\widehat{\beta}_3)$	$\widehat{\beta}_4$	$\widehat{\gamma}$
Full lik.	1000	4.00	1.00	0.16	2.00	0.20	4.00	1.00	0.16	2.00	0.00	0.01	0.14	0.20
	500	4.00	1.01	0.22	1.99	0.20	4.00	1.01	0.22	1.99	0.00	0.02	0.14	0.20
	200	4.00	1.00	0.34	1.99	0.20	4.00	1.00	0.34	1.99	0.00	0.02	0.14	0.20
	100	4.00	1.00	0.47	1.99	0.20	4.00	1.01	0.48	1.99	0.00	0.02	0.14	0.20
Cond. lik.	500	4.00	1.01	0.22	1.99	0.20	4.00	1.01	0.22	1.99	0.00	0.02	0.14	0.20
	200	4.01	1.00	0.35	1.99	0.20	4.01	1.00	0.35	1.99	0.00	0.03	0.13	0.20
	100	4.00	1.01	0.50	1.98	0.20	4.00	1.01	0.51	1.98	0.00	0.05	0.13	0.20
Gamma data	n	Normal model					NPQM model							
		$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\text{SE}(\widehat{\beta}_1)$	$\widehat{\beta}_2$	$\widehat{\gamma}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\text{SE}(\widehat{\beta}_1)$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\text{SE}(\widehat{\beta}_3)$	$\widehat{\beta}_4$	$\widehat{\gamma}$
Full lik.	1000	4.00	1.01	0.16	1.99	0.20	3.99	1.00	0.13	2.01	0.15	0.01	0.14	0.20
	500	3.97	1.15	0.24	1.98	0.20	3.97	1.08	0.18	2.00	0.15	0.02	0.14	0.20
	200	3.83	2.01	0.36	1.86	0.19	3.92	1.28	0.23	1.97	0.16	0.02	0.14	0.21
	100	3.72	2.95	0.26	1.70	0.16	3.87	1.44	0.26	1.94	0.16	0.02	0.14	0.23
Cond. lik.	500	4.00	1.00	0.22	1.99	0.20	4.00	1.00	0.19	2.01	0.15	0.02	0.14	0.20
	200	4.00	0.99	0.35	1.98	0.20	4.01	1.00	0.29	2.00	0.15	0.03	0.14	0.20
	100	4.00	0.99	0.50	1.97	0.20	4.01	1.01	0.41	2.00	0.15	0.05	0.13	0.20

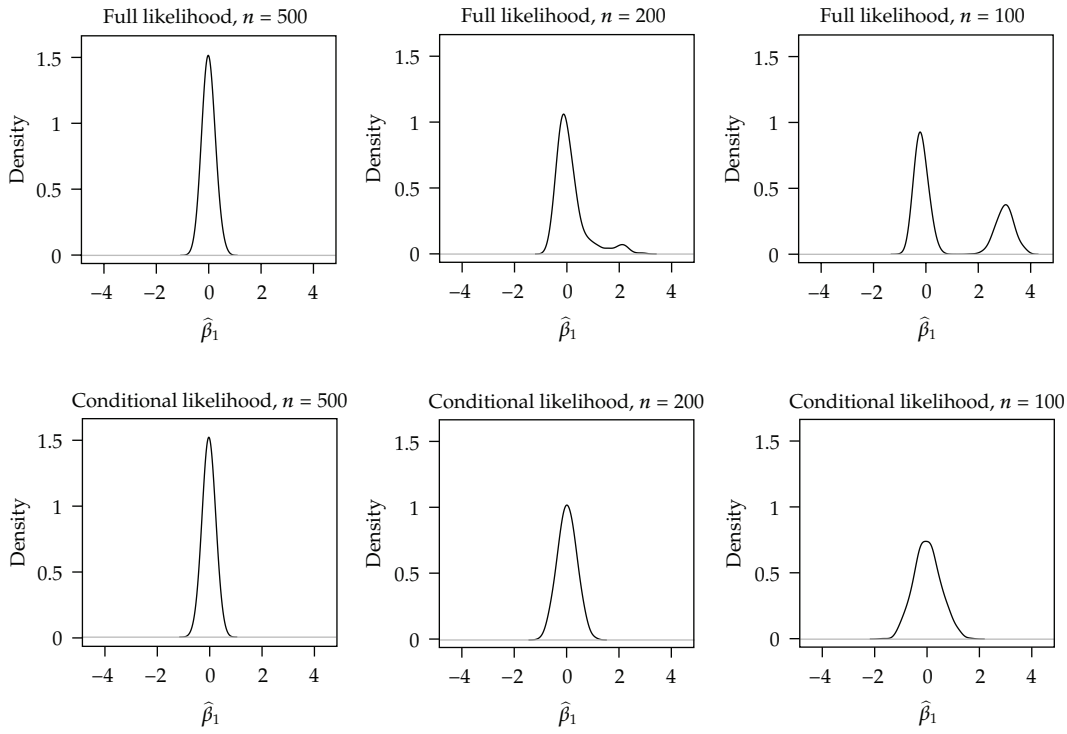

Figure 3: Sampling distributions for maximum likelihood estimates of regression coefficient β_1 from 1000 replications; normal model fitted to gamma data ($Y_i \sim \text{Gamma}(4,1)$).

sample unavailability or unsuccessful genotyping, the genotype for the lactase persistence SNP rs4988235 was unavailable for 156 individuals, giving $|\mathcal{O} \setminus \mathcal{M}| = 1660$. The model for the data described in Section 1 was specified as follows: Hardy-Weinberg model $Z_i \mid \gamma \sim \text{Binom}(2,\gamma)$ was used for the number of lactase persistence alleles. The model for BMI, which is the response variable of interest, given age at baseline $X_i \equiv b_i$ and the genotype Z_i, was taken to be $Y_i \mid b_i, Z_i, \beta \sim \text{NPQM}(\beta_0 + \beta_1 b_i + \beta_2 \mathbf{1}_{\{Z_i=0\}}, \beta_3, \beta_4, \beta_5)$. The model for the survival outcome was taken to be $P(T_i \in dt_i, E_i = e_i \mid Y_i, Z_i, \alpha) \overset{\alpha}{\propto} [\lambda_i(t_i)]^{\mathbf{1}_{\{e_i=1\}}} S_i(t_i)$, where $S_i(t_i) \equiv \exp\{\int_0^{t_i} \lambda_i(u)du\}$, and using again the Weibull form for the hazard function $\lambda_i(u) \equiv (\alpha_4/\alpha_0)(u/\alpha_0)^{\alpha_4-1} \exp\{\alpha_1 b_i + \alpha_2 Y_i + \alpha_3 \mathbf{1}_{\{Z_i=0\}}\}$. In both regression models, the lactase persistence allele noncarriers are compared to the hetero- and homozygote carriers of the allele. Under the case-cohort design where the subcohort sampling probabilities depend on b_i, the conditional likelihood expression takes the form

$$P\left(T_{\mathcal{O}\setminus\mathcal{M}}, E_{\mathcal{O}\setminus\mathcal{M}}, Y_{\mathcal{O}\setminus\mathcal{M}} \mid R_C, M_{\mathcal{O}}, T_{C\setminus\mathcal{O}\cup\mathcal{M}}, E_{C\setminus\mathcal{O}\cup\mathcal{M}}, Y_{C\setminus\mathcal{O}\cup\mathcal{M}}, X_C, Z_{\mathcal{O}\setminus\mathcal{M}}, \theta\right)$$
$$\overset{(\alpha,\beta)}{\propto} \prod_{i\in\mathcal{O}\setminus\mathcal{M}} \frac{P(T_i, E_i \mid Y_i, Z_i, \alpha) P\left(Y_i \mid b_i, Z_i, \beta\right)}{\int_{y_i} [1-(1-\pi(b_i))S_i(c_i)] P\left(dy_i \mid b_i, Z_i, \beta\right)}, \tag{6.2}$$

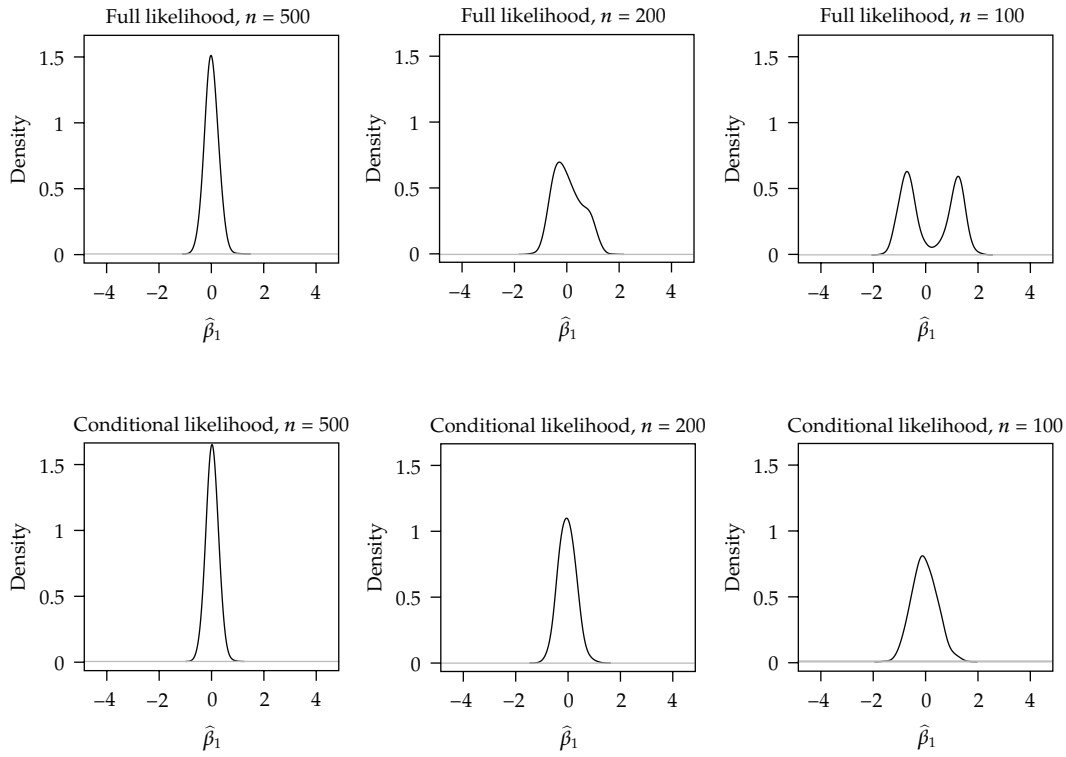

Figure 4: Sampling distributions for maximum likelihood estimates of regression coefficient β_1 from 1000 replications; NPQM model fitted to gamma data ($Y_i \sim \text{Gamma}(4,1)$).

or

$$
\begin{aligned}
&P\left(M_{\mathcal{O}}, T_{\mathcal{O}}, E_{\mathcal{O}}, Y_{\mathcal{O}}, Z_{\mathcal{O}\setminus\mathcal{M}} \mid R_{\mathcal{C}}, T_{\mathcal{C}\setminus\mathcal{O}}, E_{\mathcal{C}\setminus\mathcal{O}}, Y_{\mathcal{C}\setminus\mathcal{O}}, X_{\mathcal{C}}, \theta\right) \\
&\overset{(\alpha,\beta,\gamma)}{\propto} \prod_{i\in\mathcal{O}\setminus\mathcal{M}} \frac{P(T_i, E_i \mid Y_i, Z_i, \alpha) P\left(Y_i \mid b_i, Z_i, \beta\right) P\left(Z_i = z_i \mid \gamma\right)}{\sum_{z_i} \int_{y_i} [1 - (1 - \pi(b_i)) S_i(c_i)] P\left(dy_i \mid b_i, z_i, \beta\right) P\left(Z_i = z_i \mid \gamma\right)} \\
&\quad \times \prod_{i\in\mathcal{M}} \frac{\sum_{z_i} P(T_i, E_i \mid Y_i, z_i, \alpha) P\left(Y_i \mid b_i, z_i, \beta\right) P\left(Z_i = z_i \mid \gamma\right)}{\sum_{z_i} \int_{y_i} [1 - (1 - \pi(b_i)) S_i(c_i)] P\left(dy_i \mid b_i, z_i, \beta\right) P\left(Z_i = z_i \mid \gamma\right)},
\end{aligned}
\tag{6.3}
$$

depending on whether we condition upon the observed genotype data $Z_{\mathcal{O}\setminus\mathcal{M}}$, or whether this is modeled as part of the likelihood (see Section 5.1.3). The difference between these two approaches is that the former requires more assumptions on the missingness mechanism. Therefore, we compare here the two approaches to see whether there is any observable difference between the results. In addition, expression (6.3) enables estimation of the population allele frequency γ. However, it should be noted that while the summation over the single SNP genotype variable Z_i is not a computational problem in the present example, this is not necessarily the case generally, with multiple continuous covariates in the model. Therefore,

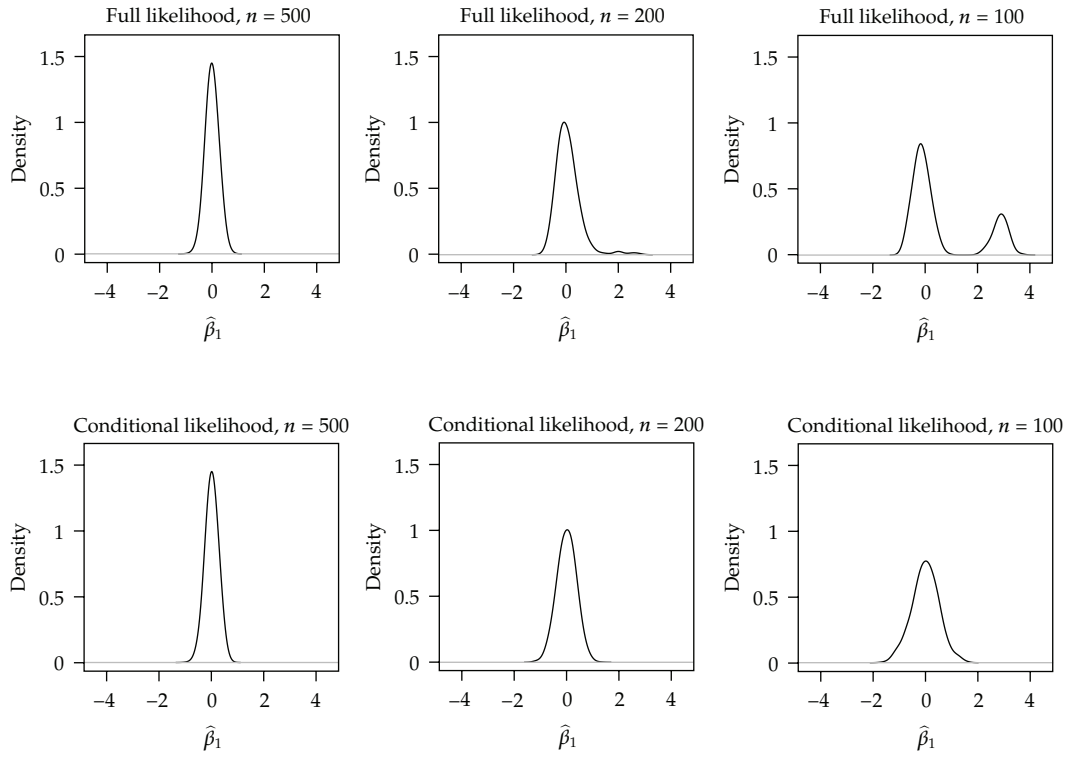

Figure 5: Sampling distributions for maximum likelihood estimates of regression coefficient β_1 from 1000 replications; normal model fitted to NPQM data ($Y_i \sim \text{NPQM}(4, 2, 0.15, 0.14)$).

the potential advantage of expression such as (6.2) is that it avoids specification of the covariate distribution and the integration over the missing covariate values.

A remaining issue to be considered is the numerical evaluation of the integrals over y_i in the correction terms of (6.2) and (6.3). It should be noted that these have to be evaluated at each parameter value tried at the hill-climbing numerical optimization. Although numerical (quadrature) integration would be feasible in one dimension, here we propose evaluation of such integrals using Monte Carlo integration by repeated sampling from the distribution $P(Y_i \in dy_i \mid b_i, Z_i, \beta)$. This has an added advantage that while the numerical evaluation of the density of the NPQM model is computationally expensive, random variates can be easily simulated from this distribution since it is defined through the quantile function. This proceeds by drawing random variates from $[0, 1]$-uniform distribution and applying formula (6.2) of Karvanen [20]. With random variates $y_i^{(j)} \sim \text{NPQM}(\beta_0 + \beta_1 b_i + \beta_2 \mathbf{1}_{\{Z_i=0\}}, \beta_3, \beta_4, \beta_5)$, $j = 1, \ldots, k$, drawn for each $i \in \mathcal{O} \setminus \mathcal{M}$ given the current parameter values, the integrals in (6.2) are then approximated by the means $(1/k) \sum_{j=1}^{k} [1 - (1 - \pi(b_i)) P(T_i > c_i \mid y_i^{(j)}, Z_i, \alpha)]$.

Maximum likelihood estimates obtained by numerical maximization of the expressions (6.2) and (6.3) with respect to α, β, and in the latter case also γ, are presented in Table 4 ($\widehat{\alpha}$ and $\widehat{\gamma}$) and Table 5 ($\widehat{\beta}$). For comparison, we also maximised a full likelihood of the form (5.1) with respect to (α, β, γ) and a weighted pseudolikelihood of the form (5.2) with respect to β. In the latter case, we accounted for the missingness within the case-cohort set by fitting

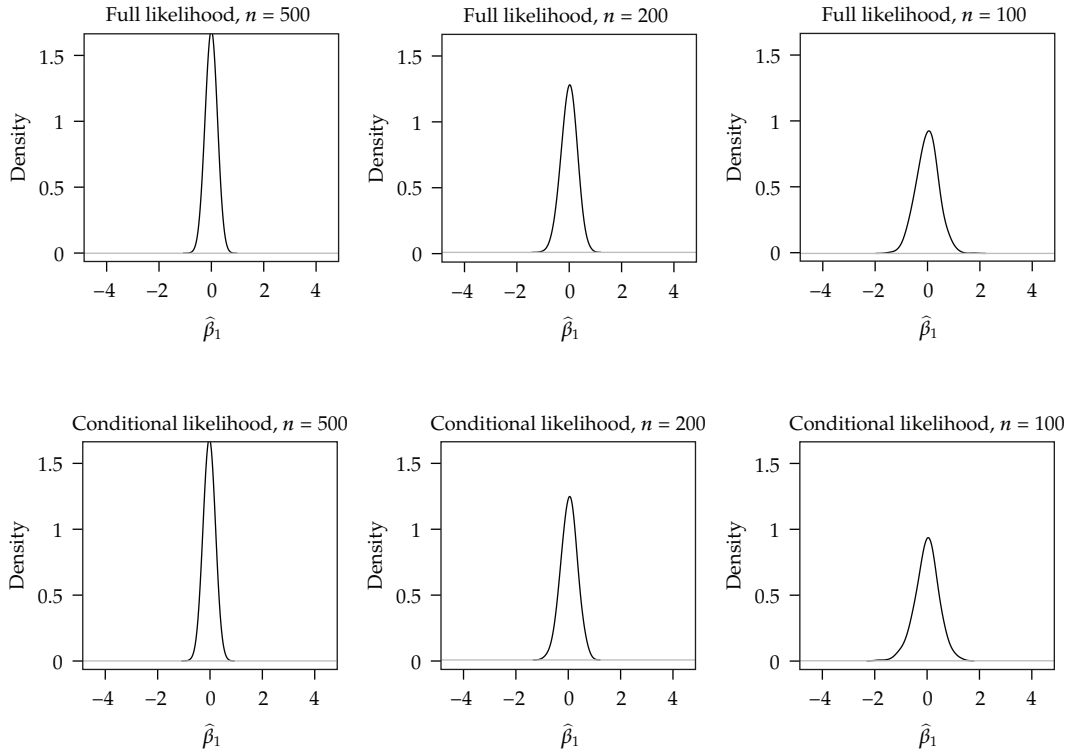

Figure 6: Sampling distributions for maximum likelihood estimates of regression coefficient β_1 from 1000 replications; NPQM model fitted to NPQM data ($Y_i \sim \text{NPQM}(4, 2, 0.15, 0.14)$).

a logistic regression model $\text{logit}\{P(M_i = 0 \mid E_i, b_i, Y_i, \eta)\} = \eta_0 + \eta_1 E_i + \eta_2 b_i + \eta_3 Y_i$ in the set $\mathcal{O}$ and calculated the adjusted weights as inverses of $P(M_i = 0 \mid E_i, b_i, Y_i, \widehat{\eta})P(R_i = 1 \mid E_i, b_i)$. Due to the extra estimation step, we used bootstrap with 5000 replications to obtain standard errors for the weighted pseudolikelihood estimates. Both conditional likelihood expressions gave very similar results, suggesting that the missing data within the case-cohort set is not a major issue in the present case. The comparison between different number of Monte Carlo replicates in numerical evaluation of the conditional likelihood correction term suggests that the estimates do no longer appreciably change when k is increased from 1000. However, the change in the estimates when increasing k from 100 to 1000 is already well within the standard errors.

Full likelihood and weighted pseudolikelihood estimates agreed well with the conditional likelihood ones, although the latter had higher standard errors, as was the case also in the simulations. As noted by Kettunen et al. [10], the absence of the lactase persistence allele shows association with lower body mass index. The residuals from the model for BMI are significantly skewed to the right, as indicated by the estimates of β_4.

7. Discussion

Although conditional logistic likelihood is well known in the context of risk set sampling designs (e.g., [5, 18, 40]), its connection to the general concept of conditional likelihood has

Table 4: Maximum likelihood estimates (standard errors) of the parameters α of the all-cause mortality model $\lambda_i(t) = (\alpha_4/\alpha_0)(t/\alpha_0)^{\alpha_4-1} \exp\{\alpha_1 b_i + \alpha_2 Y_i + \alpha_3 \mathbf{1}_{\{Z_i=0\}}\}$, and γ, the population frequency of the lactase persistence allele.

Likelihood	k	$\widehat{\alpha}_0$	$\log \widehat{\alpha}_0$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_3$	$\widehat{\alpha}_4$	$\log \widehat{\alpha}_4$	$\widehat{\gamma}$
Conditional	100	23.88	3.17 (0.21)	0.12 (0.01)	0.00 (0.01)	−0.01 (0.12)	1.34	0.29 (0.03)	—
(6.2)	1000	23.52	3.16 (0.21)	0.12 (0.01)	0.00 (0.01)	−0.01 (0.12)	1.34	0.29 (0.03)	—
	5000	23.45	3.15 (0.21)	0.12 (0.01)	0.00 (0.01)	−0.01 (0.12)	1.34	0.29 (0.03)	—
Conditional	100	20.43	3.02 (0.20)	0.12 (0.01)	−0.01 (0.01)	−0.01 (0.11)	1.33	0.29 (0.03)	0.59 (0.01)
(6.3)	1000	20.34	3.01 (0.20)	0.12 (0.01)	−0.01 (0.01)	−0.01 (0.11)	1.33	0.29 (0.03)	0.59 (0.01)
	5000	20.44	3.02 (0.20)	0.12 (0.01)	−0.01 (0.01)	−0.01 (0.11)	1.33	0.29 (0.03)	0.59 (0.01)
Full	—	18.40	2.91 (0.17)	0.09 (0.01)	−0.01 (0.01)	−0.07 (0.11)	1.33	0.28 (0.03)	0.59 (0.01)

Table 5: Maximum likelihood estimates (standard errors) of the parameters β of the BMI model $Y_i \mid b_i, Z_i, \beta \sim \mathrm{NPQM}(\beta_0 + \beta_1 b_i + \beta_2 \mathbf{1}_{\{Z_i=0\}}, \beta_3, \beta_4, \beta_5)$.

Likelihood	k	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\log(\widehat{\beta}_3/\sqrt{\pi})$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\log \widehat{\beta}_5$
Conditional	100	26.91 (0.14)	−0.07 (0.02)	−0.64 (0.27)	4.27	0.88 (0.02)	0.14 (0.03)	0.37	−1.00 (0.05)
(6.2)	1000	26.92 (0.14)	−0.07 (0.02)	−0.64 (0.27)	4.27	0.88 (0.02)	0.14 (0.03)	0.37	−1.00 (0.05)
	5000	26.92 (0.14)	−0.07 (0.02)	−0.64 (0.27)	4.27	0.88 (0.02)	0.14 (0.03)	0.37	−1.00 (0.05)
Conditional	100	26.93 (0.13)	−0.07 (0.02)	−0.63 (0.26)	4.25	0.87 (0.02)	0.14 (0.03)	0.37	−1.00 (0.05)
(6.3)	1000	26.93 (0.13)	−0.07 (0.02)	−0.63 (0.26)	4.25	0.87 (0.02)	0.14 (0.03)	0.37	−1.00 (0.05)
	5000	26.93 (0.13)	−0.07 (0.02)	−0.63 (0.26)	4.25	0.87 (0.02)	0.14 (0.03)	0.37	−1.00 (0.05)
Full	—	26.94 (0.07)	−0.07 (0.01)	−0.60 (0.25)	4.06	0.83 (0.01)	0.13 (0.01)	0.35	−1.06 (0.03)
Weighted	—	26.94 (0.15)	−0.06 (0.03)	−0.68 (0.33)	4.30	0.88 (0.04)	0.17 (0.05)	0.41	−0.89 (0.11)

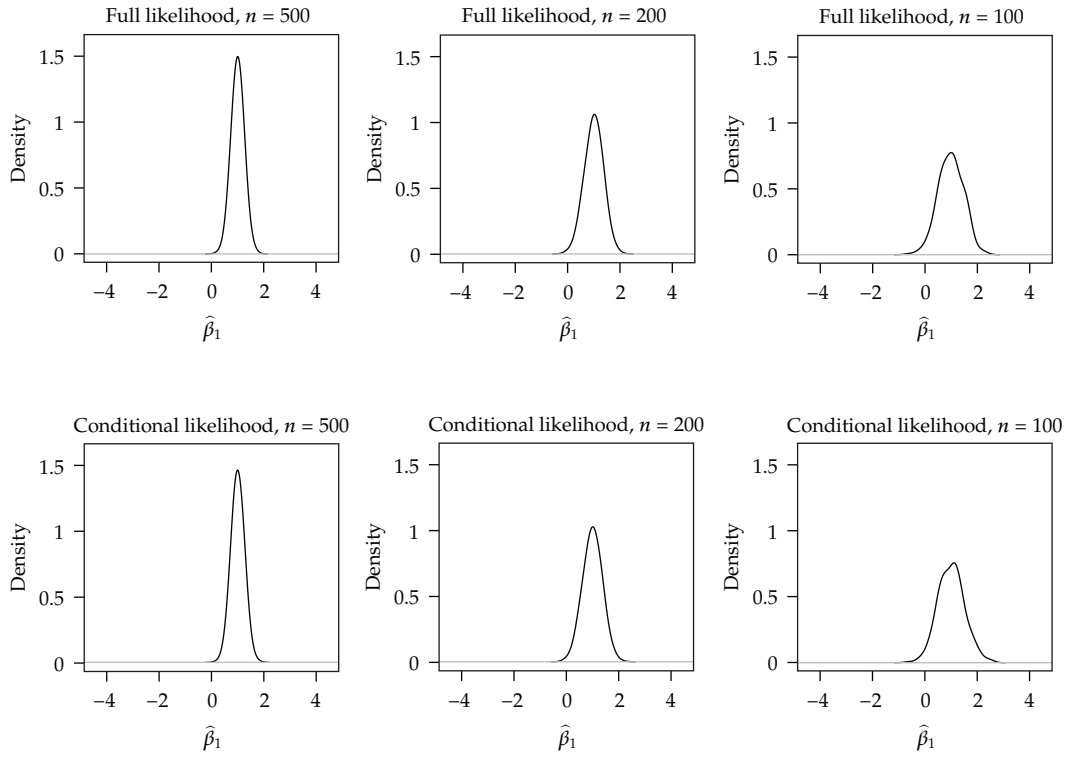

Figure 7: Sampling distributions for maximum likelihood estimates of regression coefficient β_1 from 1000 replications; normal model fitted to normal data ($Y_i \sim Z_i + N(4,4)$).

rarely been emphasized by authors. Furthermore, use of the conditional likelihood approach is not limited to the binary event outcome response situation. In this paper we have attempted to highlight these issues in the context of modeling a continuous secondary response variable under cohort sampling designs. Naturally, the conditional likelihood approach is also valid for the primary time-to-event outcome. Although here we considered only parametric specifications for the time-to-event outcome, semiparametric maximum likelihood techniques could also be utilized here. A potential disadvantage of the likelihood-based methods for secondary analysis is that they require specification of a model for the primary time-to-event outcome even though it is not of main interest in the secondary analysis setting, whereas specification of this model is avoided in the weighted pseudolikelihood approach.

Compared to the full likelihood approach which would be applicable under any cohort sampling design, the advantage of the conditional likelihood is that modeling of the population distribution of the covariates collected in the second phase can be avoided, if this is not of primary interest. In addition, as we demonstrated in a simulation example, full likelihood expressions with most of the covariate data unobserved may no longer be well behaved, a problem which the conditional likelihood approach does not have. The disadvantage of the conditional likelihood approach is that the second-phase sampling mechanism, that is, the joint distribution of the inclusion indicators, needs to be specified. In the case-cohort/Bernoulli sampling situation, this is straightforward, but in mechanisms such as risk set sampling, where the sampling probabilities are specified only implicitly, resolving the joint

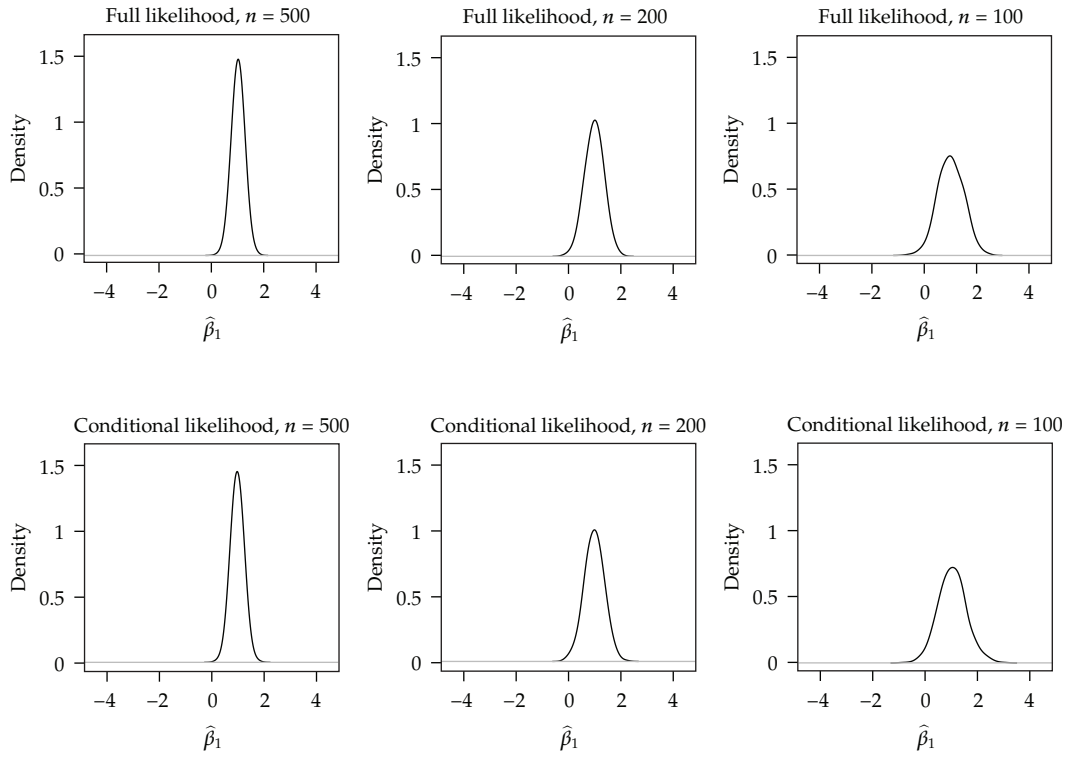

Figure 8: Sampling distributions for maximum likelihood estimates of regression coefficient β_1 from 1000 replications; NPQM model fitted to normal data ($Y_i \sim Z_i + N(4,4)$).

sampling probability can be computationally nontrivial, and approximations may be needed in practice. This is avoided in the full likelihood approach, since it does not require specification of the sampling mechanism. The inverse-probability-weighted estimating function only requires specification of the first-order selection probabilities, but the possible dependencies induced by the sampling mechanism need to be addressed in the variance estimation step. The full and conditional likelihood approaches gave equivalent efficiencies in our simulated setting, although the result might be different if the first-phase data involves covariates which are highly predictive of the second-phase covariate of interest. In any case, both likelihood-based methods gave a clear improvement in efficiency in the secondary analysis setting compared to the inverse-probability-weighted method.

Appendices

A. On Inverse-Probability-Weighted Pseudolikelihood Estimators

The use of the expression (3.1) in approximating the corresponding complete data log-likelihood $\sum_{i\in C} \log P(Y_i \mid X_i, Z_i, \beta)$, where we denote $\log P(Y_i \mid X_i, Z_i, \beta) \equiv l_i(\beta)$, is justified by considering the expectation of the pseudoscore function with respect to the joint distribution $P(R_C, T_C, E_C, Y_C, Z_C, X_C) = P(R_C \mid T_C, E_C, X_C, Y_C)P(T_C, E_C, Y_C, Z_C, X_C)$, which is assumed to

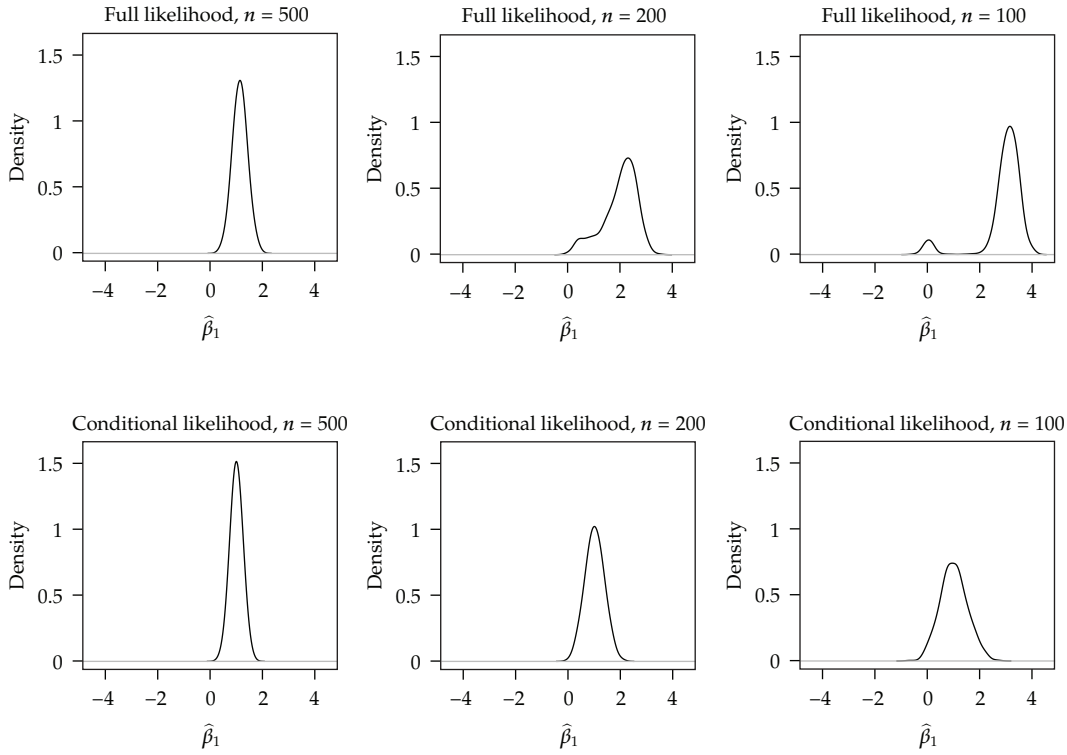

Figure 9: Sampling distributions for maximum likelihood estimates of regression coefficient β_1 from 1000 replications; normal model fitted to gamma data ($Y_i \sim Z_i + \text{Gamma}(4,1)$).

be the correct data generating mechanism. This becomes

$$
\begin{aligned}
&E_{R_C,T_C,E_C,X_C,Y_C,Z_C}\left[\sum_{i\in C}\frac{\mathbf{1}_{\{R_i=1\}}\partial \log P(Y_i \mid X_i, Z_i, \beta)/\partial\beta}{P(R_i = 1 \mid T_C, E_C, X_C, Y_C)}\right] \\
&\quad = \int_{t_C}\int_{x_C}\int_{y_C}\int_{z_C}\sum_{e_C}\sum_{i\in C}\sum_{r_i}\frac{\mathbf{1}_{\{r_i=1\}}l_i'(\beta)\sum_{r_{C\setminus i}}P(R_C = r_C \mid t_C, e_C, x_C, y_C)}{P(R_i = 1 \mid t_C, e_C, x_C, y_C)} \\
&\qquad \times P(T_C \in dt_C, E_C = e_C, X_C \in dx_C, Y_C \in dy_C, Z_C \in dz_C) \\
&\quad = \sum_{i\in C}\int_{x_i}\int_{y_i}\int_{z_i} l_i'(\beta)\int_{t_i}\sum_{e_i}P(T_i \in dt_i, E_i = e_i \mid x_i, y_i, z_i) \\
&\qquad \times P(X_i \in dx_i, Y_i \in dy_i, Z_i \in dz_i) = \sum_{i\in C}E_{X_i,Y_i,Z_i}\left[l_i'(\beta)\right] \\
&\quad = \sum_{i\in C}E_{X_i,Z_i}\left\{E_{Y_i|X_i,Z_i}\left[\frac{\partial \log P(Y_i \mid X_i, Z_i, \beta)}{\partial\beta}\right]\right\},
\end{aligned}
\tag{A.1}
$$

the last form being the expectation of the complete data score function for parameters β, which becomes zero through the inner expectation if the parametric model $P(Y_i \mid X_i, Z_i, \beta)$

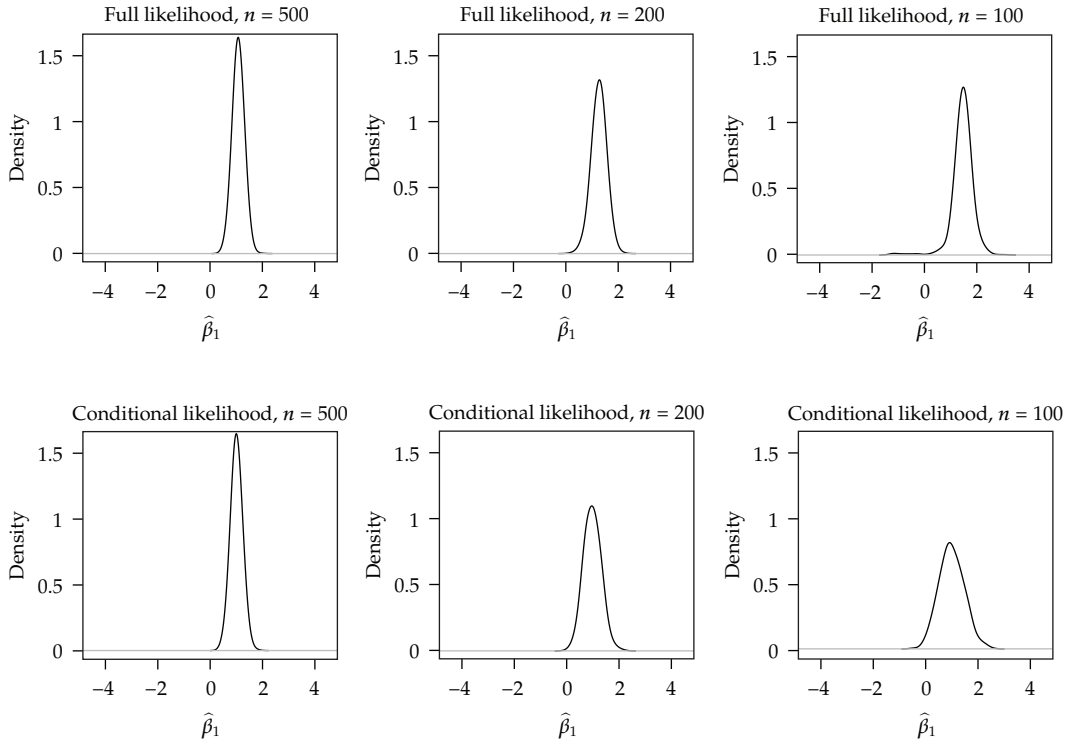

Figure 10: Sampling distributions for maximum likelihood estimates of regression coefficient β_1 from 1000 replications; NPQM model fitted to gamma data ($Y_i \sim Z_i + \text{Gamma}(4, 1)$).

is correct, that is, corresponds to the data generating mechanism. Thus, the expression (3.1) gives a valid estimating equation for parameters β. However, since the variance of the pseudoscore function does not equal the Fisher information, a robust/sandwich type variance estimator must be used instead of the inverse of the observed information. As an example, below we derive this in the special case of a Bernoulli sampling mechanism $P(R_C \mid T_C, E_C, Y_C, X_C) = \prod_{i \in C} P(R_i \mid T_i, E_i, Y_i, X_i)$, where this is straightforward. For consideration of asymptotic variances of inverse-probability-weighted estimators under stratified without replacement sampling or risk set sampling, we refer to Breslow and Wellner [41] and Cai and Zheng [42].

Denoting $q(\beta) \equiv \sum_{i \in \mathcal{O}} \log P(Y_i \mid X_i, Z_i, \beta)/P(R_i \mid T_i, E_i, Y_i, X_i)$ and $\widehat{\beta} \equiv \arg\max_\beta q(\beta)$, the latter is given as a solution to the estimating equation $q'(\beta) = 0$, or approximately, by the first-order Taylor expansion at the true value β_0, as a solution to $q'(\beta_0) + q''(\beta_0)(\beta - \beta_0) = 0$, which in turn, by substituting expected to observed pseudoinformation, gives the approximate relationship $\widehat{\beta} - \beta_0 \approx E[-q''(\beta_0)]^{-1} q'(\beta_0)$. Taking covariance of both sides then gives $\operatorname{cov}\widehat{\beta} \approx E[-q''(\beta_0)]^{-1} \operatorname{cov}[q'(\beta_0)] E[-q''(\beta_0)]^{-1}$, where the pseudo-Fisher information is given by

$$\begin{aligned} E\left[-q''(\beta_0)\right] &= E\left[-\sum_{i \in C} \frac{\mathbf{1}_{\{R_i=1\}} l_i''(\beta_0)}{P(R_i = 1 \mid T_i, E_i, X_i, Y_i)}\right] \\ &= \sum_{i \in C} E_{X_i, Z_i}\left\{E_{Y_i \mid X_i, Z_i}\left[-l_i''(\beta_0)\right]\right\}. \end{aligned} \tag{A.2}$$

Since the pseudoscore function has zero mean and now $R_i \perp R_j$, $i \neq j$, its covariance is given by

$$\begin{aligned}\operatorname{cov}\left[q'(\beta_0)\right] &= E\left[\sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{C}}\frac{\mathbf{1}_{\{R_i=1\}}\mathbf{1}_{\{R_j=1\}}l_i'(\beta_0)l_j'(\beta_0)^T}{P(R_i=1\mid T_i,E_i,X_i,Y_i)P(R_j=1\mid T_j,E_j,X_j,Y_j)}\right]\\ &= \sum_{i\in\mathcal{C}}E\left[\frac{\mathbf{1}_{\{R_i=1\}}l_i'(\beta_0)l_i'(\beta_0)^T}{P(R_i=1\mid T_i,E_i,X_i,Y_i)^2}\right]\\ &= \sum_{i\in\mathcal{C}}E_{T_i,E_i,X_i,Y_i,Z_i}\left[\frac{l_i'(\beta_0)l_i'(\beta_0)^T}{P(R_i=1\mid T_i,E_i,X_i,Y_i)}\right].\end{aligned} \tag{A.3}$$

In practice, these quantities are evaluated at the maximum likelihood point, replacing the expected quantities by their observed counterparts as

$$E\left[-q''(\beta_0)\right] \approx -\sum_{i\in\mathcal{O}}\frac{l_i''\left(\widehat{\beta}\right)}{P(R_i=1\mid T_i,E_i,X_i,Y_i)}, \tag{A.4}$$

$$\operatorname{cov}\left[q'(\beta_0)\right] \approx \sum_{i\in\mathcal{O}}\frac{l_i'\left(\widehat{\beta}\right)l_i'\left(\widehat{\beta}\right)^T}{P(R_i=1\mid T_i,E_i,X_i,Y_i)^2}. \tag{A.5}$$

Alternatively, bootstrap variance estimation might be utilized, but it should be noted that standard bootstrap would be valid only under Bernoulli type sampling designs, which can be interpreted as sampling from infinite population, whereas dependencies induced by sampling without replacement would require application of finite population bootstrap methods (e.g., [43, 44, page 37]).

B. Relationship to Retrospective Likelihood

Consider a special case where the observed data are only $(R_{\mathcal{C}}, E_{\mathcal{C}}, Y_{\mathcal{O}}, Z_{\mathcal{O}})$, where E_i, $i \in \mathcal{C}$, are binary indicators for the case status. This corresponds to the case-control situation considered by Jiang et al. [6] and Lin and Zeng [7]. The set $\mathcal{C}$ now represents a study base from which the cases and controls have been selected. Now choosing a partitioning $W = (Y_{\mathcal{O}}, Z_{\mathcal{O}})$ and $V = (R_{\mathcal{C}}, E_{\mathcal{C}})$, and supposing only the case status information has been used in the selection of cases and controls, we obtain a conditional likelihood

$$\begin{aligned}P(Y_{\mathcal{O}}, Z_{\mathcal{O}} \mid R_{\mathcal{C}}, E_{\mathcal{C}}, \theta) &= \frac{P(R_{\mathcal{C}} \mid E_{\mathcal{C}})P(E_{\mathcal{C}}, Y_{\mathcal{O}}, Z_{\mathcal{O}} \mid \theta)}{P(R_{\mathcal{C}} \mid E_{\mathcal{C}})P(E_{\mathcal{C}} \mid \theta)}\\ &= \prod_{i\in\mathcal{O}}\frac{P(E_i \mid Y_i, Z_i, \alpha)P(Y_i \mid Z_i, \beta)P(Z_i \mid \gamma)}{\int_{y_i}\int_{z_i}P(E_i \mid y_i, z_i, \alpha)P(Y_i \in dy_i \mid z_i, \beta)P(Z_i \in dz_i \mid \gamma)}.\end{aligned} \tag{B.1}$$

The last form is the retrospective form suggested by Lin and Zeng [7, page 257] for secondary analysis under case-control designs. Thus, retrospective likelihood is a special case of conditional likelihood. The advantage of the above expression is that the terms specifying the sampling mechanism cancel out. The drawbacks are that the baseline risk level (part of α) is not identifiable from this expression (see, e.g., [30, page 1076]) and that the parameters γ have to be estimated or otherwise made to disappear (Lin and Zeng [7] applied profile likelihood). However, nothing prevents us from choosing a more useful partitioning of the observed data. For instance, with $W = Y_{\mathcal{O}}$ and $V = (R_{\mathcal{C}}, E_{\mathcal{C}}, Z_{\mathcal{O}})$, we obtain

$$\begin{aligned} P(Y_{\mathcal{O}} \mid R_{\mathcal{C}}, E_{\mathcal{C}}, Z_{\mathcal{O}}, \theta) &= \frac{P(R_{\mathcal{C}} \mid E_{\mathcal{C}}) P(E_{\mathcal{C}}, Y_{\mathcal{O}}, Z_{\mathcal{O}} \mid \theta)}{P(R_{\mathcal{C}} \mid E_{\mathcal{C}}) P(E_{\mathcal{C}}, Z_{\mathcal{O}} \mid \theta)} \\ &= \prod_{i \in \mathcal{O}} \frac{P(E_i \mid Y_i, Z_i, \alpha) P(Y_i \mid Z_i, \beta)}{\int_{y_i} P(E_i \mid y_i, Z_i, \alpha) P(Y_i \in dy_i \mid Z_i, \beta)}, \end{aligned} \tag{B.2}$$

or, with $W = (E_{\mathcal{O}}, Y_{\mathcal{O}})$ and $V = (R_{\mathcal{C}}, E_{\mathcal{C}\setminus\mathcal{O}}, Z_{\mathcal{O}})$, analogously to Section 4.2,

$$\begin{aligned} &P(E_{\mathcal{O}}, Y_{\mathcal{O}} \mid R_{\mathcal{C}}, E_{\mathcal{C}\setminus\mathcal{O}}, Z_{\mathcal{O}}, \theta) \\ &\quad \overset{\theta}{\propto} \frac{P(E_{\mathcal{C}}, Y_{\mathcal{O}}, Z_{\mathcal{O}} \mid \theta)}{P(R_{\mathcal{C}} \mid E_{\mathcal{C}\setminus\mathcal{O}}, \theta) P(E_{\mathcal{C}\setminus\mathcal{O}}, Z_{\mathcal{O}} \mid \theta)} \\ &\quad = \frac{\prod_{i \in \mathcal{O}} P(E_i \mid Y_i, Z_i, \alpha) P(Y_i \mid Z_i, \beta)}{\int_{y_i : i \in \mathcal{O}} \sum_{e_i : i \in \mathcal{O}} P(R_{\mathcal{C}} \mid E_{\mathcal{C}}) \prod_{i \in \mathcal{O}} P(E_i = e_i \mid y_i, Z_i, \alpha) P(Y_i \in dy_i \mid Z_i, \beta)}. \end{aligned} \tag{B.3}$$

The baseline risk level may be identifiable from this last expression, depending on the sampling mechanism used (cf. [18]). The misconception that case-control design necessitates the use of retrospective likelihood has been recently criticized by Langholz [45] and results from equating likelihood function to a data generating mechanism. However, the order of the data collection does not need to determine how the likelihood is factorized, as long as the sampling mechanism is properly taken into account.

C. Mean and Variance of the Conditional Likelihood Score Function

In the following, we suppress the covariates $X_{\mathcal{C}}$ from the notation since these are always conditioned upon and do not affect the derivations. Also, for notational simplicity, the parameter θ is taken to be a scalar, and the notation is compressed so that, for example, $P(dz_{\mathcal{O}} \mid \theta) \equiv P(Z_{\mathcal{O}} \in dz_{\mathcal{O}} \mid \theta) \equiv P(Z_i \in dz_i \text{ for all } i \in \mathcal{O} \mid \theta)$. The probability measures indexed with respect to θ are assumed to possess the regularity properties allowing interchanging the order of integration with respect to the random variables and differentiation with respect to θ.

The conditional likelihood function of interest is now

$$\begin{aligned} P(dq_{\mathcal{O}} \mid r_C, q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta) &= \frac{P(r_C \mid q_C, z_{\mathcal{O}}, \theta) P(dq_C, dz_{\mathcal{O}} \mid \theta)}{P(r_C \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta) P(dq_{C\setminus\mathcal{O}}, dz_{\mathcal{O}} \mid \theta)} \\ &\overset{\theta}{\propto} \frac{P(dq_{\mathcal{O}}, dq_{C\setminus\mathcal{O}} \mid z_{\mathcal{O}}, \theta) P(dz_{\mathcal{O}} \mid \theta)}{P(r_C \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta) P(dq_{C\setminus\mathcal{O}} \mid z_{\mathcal{O}}, \theta) P(dz_{\mathcal{O}} \mid \theta)} \\ &= \frac{P(dq_{\mathcal{O}} \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}{P(r_C \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}, \end{aligned} \tag{C.1}$$

and the corresponding score function becomes

$$\begin{aligned} \frac{\partial \log P(dq_{\mathcal{O}} \mid r_C, q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}{\partial\theta} &= \frac{\partial \log P(dq_{\mathcal{O}} \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}{\partial\theta} - \frac{\partial \log P(r_C \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}{\partial\theta} \\ &= \frac{\partial P(dq_{\mathcal{O}} \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(dq_{\mathcal{O}} \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)} - \frac{\partial P(r_C \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(r_C \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}. \end{aligned} \tag{C.2}$$

The expectation of the score function is considered with respect to the full model $P(r_C \mid q_C, z_C, \theta) P(dq_C, dz_C \mid \theta)$. The expectation of the first term in (C.2) becomes

$$\begin{aligned} &E\left[\frac{\partial P(dq_{\mathcal{O}} \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(dq_{\mathcal{O}} \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right] \\ &= \sum_{r_C} \int_{q_C} \int_{z_C} \frac{\partial P(dq_{\mathcal{O}} \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(dq_{\mathcal{O}} \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)} P(r_C \mid q_C, z_C, \theta) P(dq_C, dz_C \mid \theta) \\ &= \sum_{r_C} \int_{q_C} \int_{z_{\mathcal{O}}} \frac{\partial P(dq_{\mathcal{O}} \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(dq_{\mathcal{O}} \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)} \\ &\quad \times \int_{z_{C\setminus\mathcal{O}}} P(r_C \mid q_C, z_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta) P(dq_C, dz_{C\setminus\mathcal{O}}, dz_{\mathcal{O}} \mid \theta) \\ &= \sum_{r_C} \int_{q_C} \int_{z_{\mathcal{O}}} \frac{\partial P(dq_{\mathcal{O}} \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(dq_{\mathcal{O}} \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)} \\ &\quad \times P(r_C \mid q_C, z_{\mathcal{O}}, \theta) P(dq_{\mathcal{O}} \mid q_{C\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta) P(dq_{C\setminus\mathcal{O}}, dz_{\mathcal{O}} \mid \theta). \end{aligned} \tag{C.3}$$

In the above expressions, it should be noted that the index set $\mathcal{O}$ is fixed by r_C. Due to the assumption on unconfounded sampling mechanism, the term $P(r_C \mid q_C, z_{\mathcal{O}}, \theta)$ does not

depend on θ, and we get

$$
\begin{aligned}
&E\left[\frac{\partial P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right] \\
&\quad= \sum_{r_{\mathcal{C}}} \int_{q_{\mathcal{C}}} \int_{z_{\mathcal{O}}} \frac{\partial\left[P(r_{\mathcal{C}} \mid q_{\mathcal{C}}, z_{\mathcal{O}}, \theta) P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)\right]}{\partial\theta} P(dq_{\mathcal{C}\setminus\mathcal{O}}, dz_{\mathcal{O}} \mid \theta) \\
&\quad= \sum_{r_{\mathcal{C}}} \int_{q_{\mathcal{C}\setminus\mathcal{O}}} \int_{z_{\mathcal{O}}} \frac{\partial\left[\int_{q_{\mathcal{O}}} P(r_{\mathcal{C}} \mid q_{\mathcal{O}}, q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta) P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)\right]}{\partial\theta} P(dq_{\mathcal{C}\setminus\mathcal{O}}, dz_{\mathcal{O}} \mid \theta) \\
&\quad= \sum_{r_{\mathcal{C}}} \int_{q_{\mathcal{C}\setminus\mathcal{O}}} \int_{z_{\mathcal{O}}} \frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}{\partial\theta} P(dq_{\mathcal{C}\setminus\mathcal{O}}, dz_{\mathcal{O}} \mid \theta).
\end{aligned}
\tag{C.4}
$$

Similarly, the expectation of the second term in (C.2) becomes

$$
\begin{aligned}
&E\left[\frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right] \\
&\quad= \sum_{r_{\mathcal{C}}} \int_{q_{\mathcal{C}}} \int_{z_{\mathcal{C}}} \frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)} P(r_{\mathcal{C}} \mid q_{\mathcal{C}}, z_{\mathcal{C}}, \theta) P(dq_{\mathcal{C}}, dz_{\mathcal{C}} \mid \theta) \\
&\quad= \sum_{r_{\mathcal{C}}} \int_{q_{\mathcal{C}}} \int_{z_{\mathcal{O}}} \frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)} \\
&\qquad\times P(dq_{\mathcal{O}} \mid r_{\mathcal{C}}, q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta) P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta) P(dq_{\mathcal{C}\setminus\mathcal{O}}, dz_{\mathcal{O}} \mid \theta) \\
&\quad= \sum_{r_{\mathcal{C}}} \int_{q_{\mathcal{C}\setminus\mathcal{O}}} \int_{z_{\mathcal{O}}} \frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}{\partial\theta} P(dq_{\mathcal{C}\setminus\mathcal{O}}, dz_{\mathcal{O}} \mid \theta) \int_{q_{\mathcal{O}}} P(dq_{\mathcal{O}} \mid r_{\mathcal{C}}, q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta) \\
&\quad= \sum_{r_{\mathcal{C}}} \int_{q_{\mathcal{C}\setminus\mathcal{O}}} \int_{z_{\mathcal{O}}} \frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}{\partial\theta} P(dq_{\mathcal{C}\setminus\mathcal{O}}, dz_{\mathcal{O}} \mid \theta).
\end{aligned}
\tag{C.5}
$$

Thus, we have

$$
E\left[\frac{\partial P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right] = E\left[\frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right], \tag{C.6}
$$

and the conditional likelihood score function has zero expectation. The Fisher information is given by

$$\begin{aligned}
&E\left[-\frac{\partial^2 \log P(dq_{\mathcal{O}} \mid r_{\mathcal{C}}, q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}{\partial\theta^2}\right] \\
&\quad = E\left[\frac{\partial^2 \log P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}{\partial\theta^2}\right] - E\left[\frac{\partial^2 \log P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}{\partial\theta^2}\right] \\
&\quad = E\left[\frac{\partial^2 P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta^2}{P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right] - E\left[\frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right]^2 \\
&\qquad + E\left[\frac{\partial P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right]^2 - E\left[\frac{\partial^2 P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta^2}{P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right].
\end{aligned} \tag{C.7}$$

By writing open the expectations as above, it is easy to see that

$$\begin{aligned}
&E\left[\frac{\partial^2 P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta^2}{P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right] = E\left[\frac{\partial^2 P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta^2}{P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right], \\
&E\left[\frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)} \frac{\partial P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right] \\
&\quad = \sum_{r_{\mathcal{C}}} \int_{q_{\mathcal{C}\setminus\mathcal{O}}} \int_{z_{\mathcal{O}}} \frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)} \frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}{\partial\theta} P(dq_{\mathcal{C}\setminus\mathcal{O}}, dz_{\mathcal{O}} \mid \theta) \\
&\quad = E\left[\frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right]^2.
\end{aligned} \tag{C.8}$$

Thus, the Fisher information becomes

$$\begin{aligned}
&E\left[-\frac{\partial^2 \log P(dq_{\mathcal{O}} \mid r_{\mathcal{C}}, q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}{\partial\theta^2}\right] \\
&\quad = E\left[\frac{\partial P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right]^2 - E\left[\frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right]^2 \\
&\quad = E\left[\frac{\partial P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right]^2 + E\left[\frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right]^2 \\
&\qquad - 2E\left[\frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)} \frac{\partial P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right] \\
&\quad = E\left[\frac{\partial P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(dq_{\mathcal{O}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)} - \frac{\partial P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)/\partial\theta}{P(r_{\mathcal{C}} \mid q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}\right]^2 \\
&\quad = V\left[\frac{\partial \log P(dq_{\mathcal{O}} \mid r_{\mathcal{C}}, q_{\mathcal{C}\setminus\mathcal{O}}, z_{\mathcal{O}}, \theta)}{\partial\theta}\right],
\end{aligned} \tag{C.9}$$

that is, equal to the variance of the conditional likelihood score function.

Acknowledgments

The majority of this work was carried out when the first author was working at the Department of Chronic Disease Prevention of the National Institute for Health and Welfare, Helsinki, Finland. The work was partly supported by the European Commission through the Seventh Framework Programme CHANCES Project [HEALTH-F3-2010-242244]. The authors would like to thank Professor Jarmo Virtamo and Professor Markus Perola of the National Institute for Health and Welfare for permission to use the ATBC data in our illustration.

References

[1] L. L. Kupper, A. J. McMichael, and R. Spirtas, "A hybrid epidemiologic study design useful in estimating relative risk," *Journal of the American Statistical Association*, vol. 70, pp. 524–528, 1975.

[2] O. Miettinen, "Design options in epidemiologic research. An update," *Scandinavian Journal of Work, Environment and Health*, vol. 8, no. 1, pp. 7–14, 1982.

[3] R. L. Prentice, "A case-cohort design for epidemiologic cohort studies and disease prevention trials," *Biometrika*, vol. 73, no. 1, pp. 1–11, 1986.

[4] D. Oakes, "Survival times: aspects of partial likelihood," *International Statistical Review*, vol. 49, no. 3, pp. 235–264, 1981.

[5] B. Langholz and L. Goldstein, "Risk set sampling in epidemiologic cohort studies," *Statistical Science*, vol. 11, no. 1, pp. 35–53, 1996.

[6] Y. Jiang, A. J. Scott, and C. J. Wild, "Secondary analysis of case-control data," *Statistics in Medicine*, vol. 25, no. 8, pp. 1323–1339, 2006.

[7] D. Y. Lin and D. Zeng, "Proper analysis of secondary phenotype data in case-control association studies," *Genetic Epidemiology*, vol. 33, no. 3, pp. 256–265, 2009.

[8] O. Saarela, S. Kulathinal, E. Arjas, and E. Läärä, "Nested case-control data utilized for multiple outcomes: a likelihood approach and alternatives," *Statistics in Medicine*, vol. 27, no. 28, pp. 5991–6008, 2008.

[9] A. Salim, C. Hultman, P. Sparén, and M. Reilly, "Combining data from 2 nested case-control studies of overlapping cohorts to improve efficiency," *Biostatistics*, vol. 10, no. 1, pp. 70–79, 2009.

[10] J. Kettunen, K. Silander, O. Saarela et al., "European lactase persistence genotype shows evidence of association with increase in body mass index," *Human Molecular Genetics*, vol. 19, no. 6, pp. 1129–1136, 2009.

[11] O. P. Heinonen, J. K. Huttunen, D. Albanes et al., "The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance," *Annals of Epidemiology*, vol. 4, no. 1, pp. 1–10, 1994.

[12] A. Evans, V. Salomaa, S. Kulathinal et al., "MORGAM (an international pooling of cardiovascular cohorts)," *International Journal of Epidemiology*, vol. 34, no. 1, pp. 21–27, 2005.

[13] S. Kulathinal, J. Karvanen, O. Saarela, and K. Kuulasmaa, "Case-cohort design in practice—experiences from the MORGAM Project," *Epidemiologic Perspectives and Innovations*, vol. 4, article 15, 2007.

[14] A. J. Lee, L. McMurchy, and A. J. Scott, "Re-using data from case-control studies," *Statistics in Medicine*, vol. 16, no. 12, pp. 1377–1389, 1997.

[15] M. Reilly, A. Torrång, and A. Klint, "Re-use of case-control data for analysis of new outcome variables," *Statistics in Medicine*, vol. 24, no. 24, pp. 4009–4019, 2005.

[16] D. B. Richardson, P. Rzehak, J. Klenk, and S. K. Weiland, "Analyses of case-control data for additional outcomes," *Epidemiology*, vol. 18, no. 4, pp. 441–445, 2007.

[17] G. M. Monsees, R. M. Tamimi, and P. Kraft, "Genome-wide association scans for secondary traits using case-control samples," *Genetic Epidemiology*, vol. 33, no. 8, pp. 717–728, 2009.

[18] B. Langholz and L. Goldstein, "Conditional logistic analysis of casecontrol studies with complex sampling," *Biostatistics*, vol. 2, pp. 63–84, 2001.

[19] O. Saarela and S. Kulathinal, "Conditional likelihood inference in a case-cohort design: an application to haplotype analysis," *International Journal of Biostatistics*, vol. 3, no. 1, article 1, 2007.

[20] J. Karvanen, "Estimation of quantile mixtures via L-moments and trimmed L-moments," *Computational Statistics & Data Analysis*, vol. 51, no. 2, pp. 947–959, 2006.
[21] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York, NY, USA, 1987.
[22] J. D. Kalbfleisch and J. F. Lawless, "Likelihood analysis of multi-state models for disease incidence and mortality," *Statistics in Medicine*, vol. 7, no. 1-2, pp. 149–160, 1988.
[23] S. O. Samuelsen, "A pseudolikelihood approach to analysis of nested case-control studies," *Biometrika*, vol. 84, no. 2, pp. 379–394, 1997.
[24] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, vol. 47, pp. 663–685, 1952.
[25] S. O. Samuelsen, H. Ånestad, and A. Skrondal, "Stratified case-cohort analysis of general cohort sampling designs," *Scandinavian Journal of Statistics*, vol. 34, no. 1, pp. 103–119, 2007.
[26] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*, Chapman and Hall, London, UK, 1974.
[27] D. R. Cox, "Partial likelihood," *Biometrika*, vol. 62, no. 2, pp. 269–276, 1975.
[28] E. B. Andersen, "Asymptotic properties of conditional maximum-likelihood estimators," *Journal of the Royal Statistical Society B*, vol. 32, pp. 283–301, 1970.
[29] J. D. Kalbfleisch and D. A. Sprott, "Application of likelihood methods to models involving large numbers of parameters," *Journal of the Royal Statistical Society B*, vol. 32, pp. 175–208, 1970.
[30] S. R. Seaman and S. Richardson, "Bayesian analysis of case-control studies with categorical covariates," *Biometrika*, vol. 88, no. 4, pp. 1073–1088, 2001.
[31] D. R. Cox, *Principles of Statistical Inference*, Cambridge University Press, Cambridge, UK, 2006.
[32] J. Ma, C. I. Amos, and E. W. Daw, "Ascertainment correction for Markov chain Monte Carlo segregation and linkage analysis of a quantitative trait," *Genetic Epidemiology*, vol. 31, no. 6, pp. 594–604, 2007.
[33] S. Kim and V. De Gruttola, "Strategies for cohort sampling under the Cox proportional hazards model, application to an AIDS clinical trial," *Lifetime Data Analysis*, vol. 5, no. 2, pp. 149–172, 1999.
[34] C.-E. Särndal, B. Swensson, and J. Wretman, *Model Assisted Survey Sampling*, Springer, New York, NY, USA, 1992.
[35] S. Kulathinal and E. Arjas, "Bayesian inference from case-cohort data with multiple end-points," *Scandinavian Journal of Statistics*, vol. 33, no. 1, pp. 25–36, 2006.
[36] J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, Hoboken, NJ, USA, 2nd edition, 2002.
[37] O. Saarela, S. Kulathinal, and J. Karvanen, "Joint analysis of prevalence and incidence data using conditional likelihood," *Biostatistics*, vol. 10, no. 3, pp. 575–587, 2009.
[38] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011, http://www.r-project.org/.
[39] M. Galassi, J. Davies, J. Theiler et al., *GNU Scientific Library Reference Manual*, Network Theory Ltd., Bristol, UK, 3rd edition, 2009, http://www.gnu.org/software/gsl/.
[40] B. Langholz, "Use of cohort information in the design and analysis of case-control studies," *Scandinavian Journal of Statistics*, vol. 34, no. 1, pp. 120–136, 2007.
[41] N. E. Breslow and J. A. Wellner, "Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression," *Scandinavian Journal of Statistics*, vol. 34, no. 1, pp. 86–102, 2007.
[42] T. Cai and Y. Zheng, "Evaluating prognostic accuracy of biomarkers in nested case-control studies," *Biostatistics*, vol. 13, no. 1, pp. 89–100, 2012.
[43] J. G. Booth, R. W. Butler, and P. Hall, "Bootstrap methods for finite populations," *Journal of the American Statistical Association*, vol. 89, no. 428, pp. 1282–1289, 1994.
[44] R. J. Gray, "Weighted analyses for cohort sampling designs," *Lifetime Data Analysis*, vol. 15, no. 1, pp. 24–40, 2009.
[45] B. Langholz, "Case-control studies = odds ratios: blame the retrospective model," *Epidemiology*, vol. 21, no. 1, pp. 10–12, 2010.