Ascertainment bias in rate ratio estimation from case-sibling control studies of variable age-at-onset diseases

Bryan Langholz[1]
Argyrios Ziogas[2]
Duncan C. Thomas[1]
Cheryl Faucett[3]
Mark Huberman[1]
Larry Goldstein[4]

[01] Department of Preventive Medicine, University of Southern California

[02] Department of Preventive Medicine, University of California, Irvine

[03] Department of Biostatistics, University of California, Los Angeles

[04] Department of Mathematics, University of Southern California

[0] Correspondence to B. Langholz: Department of Preventive Medicine, University of Southern California, School of Medicine, 1540 Alcazar St. CHP-220, Los Angeles, California 90033, U.S.A., Phone: (213) 342-1212, FAX: (213) 342-2349,
e-mail: langholz@hsc.usc.edu

Abstract

Motivated by a Finnish case-control study of early onset diabetes in which diabetic children are matched to sibling controls, we investigate ascertainment bias of the usual rate ratio estimator from case-control data under simplex complete ascertainment of families during a fixed interval of time. Analytic results indicate that the assumptions necessary for valid estimation are that 1) the disease is rare and 2) the factors under study are exchangeable, essentially that the covariate distribution does not depend on calendar time or birth order. Further, we found that the rare disease assumption could be dropped by restricting to cases that were diagnosed during the enrollment period of the study or including all cases but eliminating the proband as a control for non-enrollment period cases. An important consequence of this work is that standard family-based case-control studies are subject to ascertainment bias if exchangeability of the covariates under investigation does not hold.

Some key words: Case-control studies; Censoring; Diabetes; Epidemiology; Family studies; Genetics; Partial likelihood; Proband elimination.

1 Introduction

The growing interest in genetic determinants of variable age-of-onset diseases such as cancer, diabetes, and psychiatric disorders has engendered collaborations between geneticists and epidemiologists. Statisticians are playing a key role in the emerging field of genetic epidemiology by developing study designs and appropriate methods of analysis that link the quite different approaches used by these two fields. It is increasingly common to rely on family case-control study designs for quantifying gene-disease associations and gene by environment interactions both to control for common family environment and potential "population stratification" [Curtis, 1997]. In this design, non-diseased family members serve as controls for the diseased case(s) in the family in a standard case-control study fashion.

One aspect of studies that use family controls that needs to be addressed is the enrollment of families into the study through affected probands. Because the high risk genotype is likely to be over-represented in families with a diseased member, a disportionate number of family members will be discovered to be diseased. In classical segregation analysis, this problem of "ascertainment bias" has long been recognized [Fisher, 1934; Morton, 1959] and appropriate ways to incorporate the additional familial cases have been developed. In this paper, we explore this issue in the context of a population based family study of childhood insulin dependent diabetes mellitus (IDDM) with the goal of estimating the rate ratio associated with measured HLA genotypes which we now describe.

The nationwide Childhood Diabetes in Finland (DiMe) Study of early onset IDDM coordinated at the Finnish National Public Health Institute has was created to assess genetic and environmental determinants of the disease that could explain the extraordinarily high rate of IDDM in Finland. Between September 1986 and April 1989, (which we will call the "enrollment period,") incident cases of IDDM in children up to age 15 (the "probands,") were identified through the prospective IDDM registry and their families were asked to participate in the study [Tuomilehto *et al.*, 1992]. Ninety five percent of those contacted did participate and, among other data collected, blood samples were taken from the proband's first degree relatives and serologically determined HLA haplotypes were inferred from the family data [Tuomilehto-Wolf *et al.*, 1989]. Since the HLA genotype does not change with age or the occurrence of diabetes and mortality for this age group is very low, HLA information was available for virtually all children in each ascertained sibship. Further, any additional cases of IDDM in the ascertained families diagnosed prior to February 1986, the "pre-enrollment period," or in the "post-enrollment period" from May 1989 to August 1993 were noted in the data set. The ascertainment scheme used in the DiMe study is an example of simplex complete ascertainment [Thompson, 1993] over the enrollment period. *Simplex* refers to the requirement that a single proband is sufficient to ascertain the family. *Complete* means that all families who meet this requirement are included in the study.

In this paper, we investigate the potential biases that can arise when families are ascertained through a proband and the circumstances in which it is acceptable to extend observation time outside the enrollment period. We further develop methods to correct for this bias. After describing the DiMe data as a family stratified cohort study and showing why standard methods of analysis, which restrict the follow-up to the enrollment period, lead to an undesirable loss of data, in Section 3 we describe our basic analytic approach to ascertainment bias. In Section 3.2 we show how using non-enrollment period experience results in ascertainment bias (even when disease is rare) in two-sibling families. The cause of this bias is investigated analytically and we describe the *exchangeability* assumption for unbiasedness. These findings suggest two analysis methods for general family data that, under exchangeability, lead to valid estimation. The first is to restrict to contributions from enrollment period cases only. The second is to additionally include non-enrollment case contributions but with the proband excluded as a control. The implications of these results for the DiMe analysis and for family case-control studies based on just the observed disease status at the end of the study period, i.e., ignoring time are discussed in Section 4. Mathematical details and an extensive computer simulation study are described in a Technical Report [Langholz and Huberman, 1999] available from the authors.

2 Methods of analysis

Figure 1 illustrates the basic data for five hypothetical sibships arising from the DiMe study. The appropriate time scale for comparing rates of IDDM is age and the disease experience is viewed from this perspective. Each horizontal line represents one child who may contract IDDM at an age, indicated by a •, or be disease-free by the end of the post-enrollment period. The tick marks (|) indicate the boundaries of enrollment period. Note that a subject who reaches age 15 disease-free before the end of the post-enrollment period is censored at age 15. The solid line denotes time that a subject is under observation during the enrollment period and the dashed line is observation time that is only available for sibships that are ascertained by the study.

We view these data as a family stratified cohort study. Thus, an appropriate hazard model for the rates of disease is of the stratified Cox form

$$\lambda(t, Z, f) = \lambda_f(t) \exp(Z\beta_0)$$

where $\lambda_f(t)$ is the baseline hazard, specific to family f, as a function of age t, Z is a vector of covariates that appropriately summarize genotype and/or environmental information, and β_0 are true log rate ratio slope parameters which are to be estimated. The analysis is based on a partial likelihood approach in which *risk sets* are formed consisting of the IDDM case and sibling controls who attain the age of the case in the course of the study and are disease free at that age [Cox, 1972]. These are the vertical "•, ∘" sets in the figure. It is the comparison of HLA genotypes in the case to that of the sibling controls that is the basis of rate ratio estimation in a "likelihood" that is the product of contributions from each family of the form

$$L = \prod_{d \in D} \frac{\exp(\beta' Z_d)}{\sum_{j \in \mathcal{R}_d} \exp(\beta' Z_j)} \tag{1}$$

where D is the set of case indices, including both proband and familial cases. The sum in the denominator is over members of the risk set $\mathcal{R}_d$ for case d with Z_j the covariate values for family member j. The estimator of β_0 is obtained by maximizing the likelihood over β.

Because a member of each sibship must have become diseased during the enrollment period, the first four families in Figure 1 are ascertained by the study and potentially can make contributions to the rate ratio estimation. In families 1 and 2, the proband's sibling attains the age of the proband in the course of the study period and the case-control pair consists of the proband as case and the sibling as control; these two situations are distinguished by whether or not the unaffected sibling attains the age of the proband during the enrollment period. In family 3, the proband has an older sibling who became diseased at a younger age than the proband's age of diagnosis. In this situation, the proband could serve as the control to the sibling case even though the proband would not contribute as a case because of the lack of an age-matched control. Family 4, with four siblings and two cases, could make two contributions with three siblings as controls for the youngest case and two siblings as controls for the proband. In family 5, the youngest and oldest siblings were under observation during the enrollment period but, since neither of them had contracted the disease during this period, the study would not detect their sibling who contracted disease prior to the enrollment period.

In our analysis of this data, we formed the risk sets for the DiMe data as in Figure 1 and used all available sets in our analysis [Langholz *et al.*, 1995]. The "All siblings" column of Table 1 gives the rate ratio estimates for DR3 and DR4 alleles from the "log-additive" model. This is a standard log linear model as in (1) with DR3 and DR4 covariates defined as 0, 1, or 2 depending on the number of occurrences of the respective alleles at the DR locus. We found that this model adequately described the DR3/DR4 effects in the data. Out of the cases who were ascertained by the study, 498 contribute informative risk sets, by which we mean that the case had at least one control (although some of these were completely exposure concordant and hence uninformative in that sense). As seen in Table 1, many of the cases had multiple sibling controls, the maximum being seven.

The problem with this approach is that data is collected in calendar time but age is the relevant time scale for IDDM incidence. The *standard* theory for the analysis of cohort data would require that the observation of the cohort members be left truncated at the ages corresponding to the beginning, and right censored at the end, of the enrollment period. This would mean that additional familial cases that occur outside the enrollment period could not be used in the analysis. Further, to be eligible as a control for a case that occurs during the enrollment period, a sibling would need to achieve the age of the case during the enrollment period. The former restriction is not too severe as, for rare diseases, the vast majority of cases will be the probands. (In the DiMe data, there were only 34 informative non-enrollment period risk sets out of the 498.) However, the latter restriction will result in many probands essentially being dropped from the study because of the lack of any sibling who meets the requirement. The consequences of these restrictions is illustrated in Figure 1 by only including "solid line" observation times in setting up the risk sets. Only families 1 and 4 would contribute informative risk sets, each with a single case-control pair. The results of restricting the DiMe data in this way are given in the "censored observation period" column of Table 1. Only 114 cases contribute to the analysis, 110 of which have just one control. Compared to the all siblings analysis, the resulting DR3 and DR4 rate ratio estimates are much more unstable, as indicated by the wide confidence intervals.

Thus, application of standard methodology in this situation results in an unacceptable loss of data and we chose to treat members of the ascertained families as if they had been under observation from birth until the end of the post-enrollment period. The results from this analysis were consistent with those using a quite different approach, using the possible four possible genotypes in offspring given the parent's genotypes [Self *et al.*, 1991], so we felt reassured about our overall conclusions. Nonetheless, the appropriateness of our analysis and the potential biases that could arise were open questions.

Another reasonable analysis of these data is to take a "grouped time" approach, ignoring the failure time structure altogether and modeling the probability (rather than the rate) of disease. Data would be organized by family with case-control status determined by disease status at the end of the study period. The logistic model then would be appropriate, using conditional logistic regression methods to accommodate the family stratification. For the analysis of cohort data generally, when the disease is rare and censoring does not depend on the covariates, the odds ratio estimates from grouped time analyses will be quite close to the rate ratio estimates from the corresponding risk set based analyses. In our framework, this approach assumes a "rare disease" and that non-diseased siblings will remain on study and non-diseased to the end of the post-enrollment period (e.g., to 15 years old in the DiMe study). Thus, returning to Figure 1, families 1 and 2 would make same likelihood contributions as in the risk set approach, while disease concordant family 3 does not contribute. Family 4 would contribute a single case-control set with two cases and two controls. Although we will focus on the risk set approach our findings will apply to the grouped time approximation as well; we will return to this point in the Discussion.

3 Analytic study of ascertainment bias

The key to our analytic approach was to derive the expected disease incidence in ascertained sibships, treating the subjects *as if* they had been observed from birth ($\tilde{\lambda}$), as a function of the expected incidence had all members of the sibship *in fact* been followed from birth (λ). As shown in the Appendix, for subject j in an ascertained sibship this is given by $\tilde{\lambda}_j(t) = \lambda_j(t)w_j(t)$ where $w_j(t)$ is the *distortion* due ascertainment. This is then used to compute the expectation of the score from likelihood (1) as a function of β. The solution to the expected score (equation (5) in the Appendix) set equal to zero is the "large sample" ascertainment biased parameter estimate. The distortion for simplex complex ascertainment is explicitly derived in the Technical Report and we found that $0 \leq w_j(t) \leq 1$ during the pre-enrollment period and $1 \leq w_j(t)$ during the enrollment period and is one during the post-enrollment period. This has the intuitive explanation that conditioning on the knowledge that at least one of the siblings will become diseased during the enrollment period decreases the chance that any particular sibling will become diseased during the pre-enrollment period and increases the chance of disease during the enrollment period, relative to a family picked without that knowledge. Once the ascertainment event has occurred, there is no additional information and the disease rates are determined by the undistorted intensity.

3.1 Twins

We now explore the effect on relative risk estimates of this distortion in hazard rates due to ascertainment. It is simplest to begin with the case of completely overlapping enrollment periods, which of course can occur only if the birth dates are identical, i.e., for twins. Since an unaffected twin is always a "valid" control for a proband (attains the age of the case during the enrollment period), this situation focuses exclusively on the question of the effect of using cases that occur outside the enrollment period. The somewhat surprising result is that pre-enrollment period cases are not at all informative for estimation of the rate ratio. This is not because they are more likely to be exposure concordant since such pairs would not contribute to the estimation of the rate ratio anyway. Rather, it is because the expected number of case-exposed and case-unexposed pre-enrollment discordant pairs is equal (for a rare disease), given that there must also be a case during the enrollment period.

Let f_0 and f_1 be the probabilities of failure in unexposed individuals during the pre-enrollment and enrollment periods, respectively, and let ϕ_0 be the relative risk. Now consider the contributions from pre-enrollment exposure discordant pairs. First, consider the probability of a case-exposed control-unexposed pair. The probability that the exposed sibling is the case during the pre-enrollment period is approximately $f_0\phi_0$. In a standard case-control study, the probability that the unexposed subject did not fail in this period is simply $1 - f_0$ but here, the control must have failed during the enrollment period in order to be the proband, with probability $(1 - f_0)f_1$. Thus, the proportion of such pre-enrollment pairs is $f_0\phi_0(1-f_0)f_1$. Similarly, the proportion of case-unexposed control-exposed pairs is $f_0(1 - f_0\phi_0)f_1\phi_0$. Now, the usual estimator of the rate ratio converges to the ratio of these two probabilities which is seen to be, for a rare disease, approximately equal to one.

For twins, a second case which occurs during the post-enrollment period is uninformative simply because there is no remaining control. The first cases (the probands) and their then-unaffected cotwins would still provide a valid estimator of ϕ_0, unless one restricted the analysis to such disease-concordant pairs. However, an examination of triplets where post-enrollment cases may have eligible controls, such additional cases also produce a bias towards the null, for essentially the same reason as the pre-enrollment cases do [Langholz and Huberman, 1999].

This example provides the intuition for why inclusion of the risk sets defined by cases outside the enrollment period will lead to bias toward the null.

3.2 Two-sibling families with non-overlapping enrollment periods

Because twins are perfectly aligned on calendar time, the previous example cannot be used to explore the use of a sibling as a control who attains the age of the case outside the enrollment period. To do this, we now consider the opposite extreme where the enrollment periods do not overlap at all on an age scale and begin with the simplest example that illustrates that bias can arise if non-enrollment period controls are used. Consider a population of two-child families in which one child is five years older than the other and suppose that the probability of disease is f up to age five and is $2f$ between ages five and ten. Further, assume that disease is rare so that we can ignore disease concordant pairs. Suppose we want to assess the effect of birth order, which we will assume is not associated with disease. Each pair is, by definition, birth order discordant so each case contributes a birth order discordant pair. Now if the entire study population were observed over ages 0-9 (this would require 15 years in calendar time), the expected proportion of first born case pairs is the probability that the first born is diseased over the 10 years or $f + 2f = 3f$. Since there is no association with birth order, there will be the same proportion of second born case pairs so the relative risk for birth order is one, as it should be.

Now, suppose that this population is observed "in calendar time" for five years starting at the birth of the younger child. Thus, the younger child is followed up to age five and the older from age five to ten. Suppose we estimate the effect of birth order on disease up to age ten in this group by using the case-sibling control pairs defined in a quite reasonable way, simply ignoring period of observation and age altogether and estimating the relative risk from the disease discordant pairs. In our framework, we have included the pre-enrollment experience (the older sibling control for the younger sibling case) and post-enrollment experience (the younger sibling control for the older sibling case). Now, in order for the family with a first born diseased case to be ascertained, the disease must have occurred during ages 5-9. Thus, the proportion of first born case pairs is $(1 - f)2f$. Similarly, second born cases must occur during ages 0-4 so that the proportion of second born case pairs is simply f. This yields a relative risk of about two, simply reflecting the difference in age between the first and second born children. This example illustrates that the use of non-enrollment period controls can lead to ascertainment bias, even under a rare disease assumption. We now describe analytic results which elucidate why bias occurred in this example and the conditions under which there is no bias.

3.2.1 Follow-up to the end of the enrollment period

We first consider follow-up only to the end of the enrollment period and will consider post-enrollment follow-up in the next section. Note that if observation were restricted to just the enrollment period, as standard methods would require, families of this type would not contribute to the analysis at all because of the non-overlapping enrollment periods. Thus, the inclusion of pre-enrollment experience is necessary in order to obtain any information from this type of sibship.

The contributing case-control pairs may be classified into four types given in Figure 2. The enrollment period corresponds to age intervals (a_1, a_2) and (a'_1, a'_2) for younger and older siblings, respectively. Type 1 is the only disease discordant pair, in which the younger sibling is the proband and the older sibling serves as the control. The remaining three types are disease concordant and thus much rarer than type 1.

Let Z_1, Z_2 be time-fixed dichotomous exposure indicators for the older and younger sibling, respectively, and let $p_{ij} = \mathrm{pr}(Z_1 = i, Z_2 = j), i, j = 0, 1$ be the probability of such an exposure pair in the "uncensored" population. In particular, Z_1 and Z_2 are not assumed to be independent or identically distributed. Then, as shown in the Technical Report, for a rare disease the maximum likelihood estimator from the standard conditional likelihood given by (1) $\exp(\hat{\beta})$ converges to $\phi^* = \phi_0 p_{01}/p_{10}$. Thus, the key requirement for ϕ^* to be unbiased is "exchangeability" in the sense that $p_{01} = p_{10}$. Intuitively, bias will arise if an older sibling has a different marginal probability of exposure than the younger because of, for example, secular trends in exposure, a birth order effect, or a maternal age effect. The birth order example given above is a rather extreme example of this since $p_{10} = 1$ and $p_{01} = 0$. If, in our example, we only included pre-enrollment experience, then $\phi^* = 0$ since only the second born would have a control. For the genetic factors measured in the DiMe study there is no evidence of a birth order effect, so that Z_1 and Z_2 are exchangeable.

Further insight is obtained by examining the expected score function under the exchangeability assumption. Let $A_0(a_1, a_2)$ be the cumulative baseline hazard, approximately the probability of disease in unexposed, between age a_1 and a_2. The expected score function is approximately proportional to

$$[\phi_0 - \phi] A_0(a_1, a_2) + [1 - \phi]\phi_0 [A_0(a_1, a_2) A_0(0, a_1) + O(A_0(a_1, a_2)^2) + A_0(a'_1, a'_2) A_0(0, a_1)] \quad (2)$$

with each of these terms corresponding to the contributions from each of the respective types of contributing pairs in Figure 2. As seen in the first term, the "discordant pairs" are unbiased in that they estimate ϕ_0. As with twins, the disease concordant pair contributions (the last three terms) estimate one *regardless of* ϕ_0 and thus add a bias towards the null. The weights associated with each of these terms are of the order of the probability of the occurrence of such pairs, so the discordant pairs are the most common.

3.2.2 Post-enrollment period observation

In the DiMe study, ascertained families were followed up long after the enrollment period ended, during which time 16 new cases of IDDM were diagnosed and many of the subjects attained the age of their diseased siblings without contracting IDDM, so were eligible to serve as controls. To get an idea of the effect of including this follow-up on the estimation of the rate ratio, we continue with our two-sibling family example with follow-up. In this situation, additional informative pairs arise because younger siblings attain the age of the older sibling proband without contracting IDDM or contract IDDM during the post-enrollment period yielding three additional types of pairs illustrated in Figure 3. Type 5 is disease discordant with the younger sibling disease free by the age of the proband. There are two disease concordant types (types 6 and 7) with the younger sibling diagnosed during the post-enrollment period but prior to the proband's age of diagnosis. Adding the contributions from these pairs to the expected score from the last section yields a biased rate ratio estimate,

$$\phi^* \approx \phi_0 \frac{p_{01}A_0(a_1, a_2) + p_{10}A_0(a_1', a_2')}{p_{10}A_0(a_1, a_2) + p_{01}A_0(a_1', a_2')}$$

so that unbiasedness could still be expected if $p_{01} = p_{10}$ or $A_0(a_1, a_2) = A_0(a_1', a_2')$, essentially that the disease rate remains constant over the age range of the study. Thus, to construct our example of ascertainment bias arising from the use of non-enrollment period controls in Section 3.2, we needed a situation in which both exchangeability does not hold (birth order) and the rates change with age. If exchangeability can be assumed then the expected score expression (2) is augmented by the addition of terms from pairs of types 5 to 7

$$[\phi_0 - \phi]A_0(a_1', a_2') + [1 - \phi]\phi_0[A_0(a_1', a_2')A_0(a_2, a_1') + O(A_0(a_1', a_2')^2)]. (3)$$

Again, the contribution from the disease discordant pairs is "unbiased" while those of the disease concordant pairs are estimating one, regardless of ϕ_0, and thus produce a bias towards the null.

3.3 General family structures

The above consideration of two-sibling families suggest that the two key assumptions needed for valid estimation from an "all siblings" analysis, as we used in the DiMe study, are the rare disease assumption (in particular that there are few non-enrollment period cases) and exchangeability of the covariate values. But, whatever non-enrollment period cases there are in the study are only creating bias so it is useful to have methods that do not rely on the rare disease assumption.

One method, the "enrollment period cases" method, suggested immediately from the last section, is to simply restrict to the contributions of enrollment period cases, but including "non-enrollment period controls." As shown in the Technical Report , this approach is unbiased for the two-sibling situation assuming exchangeability but not rare disease. Thus, returning to Figure 1, families 1 and 2 would contribute case-control pairs, family 3 would not be included because the case is pre-enrollment, while only the older case in family 4, with the two controls would be included.

Another method, starts with the enrollment period case-control set but, additionally includes pre-and post-enrollment cases *with the proband excluded as a control* (see Section 5 of the Technical Report). The theoretical framework we described in Section 3 does not apply for this method because exclusion of the proband as a control is not a predictable process. Thus, our insights into the assumptions which are required to ensure unbiasedness are more limited. However, simulation studies, described in the Technical Report, indicate that this method does indeed result in valid estimation under the fairly wide range of situations we examined. This method has some intuitive basis from proband elimination ascertainment correction methods [Eland-Johnson, 1971] in that the remaining siblings are representative of the source population of *familial* cases, whereas the proband's source population is the general population. Applying this approach to the families in Figure 1, families 1 and 2 each contribute a case-control pair, family 3 would be dropped because the proband cannot serve as a control for the pre-enrollment case and family 4 contributes two case-control sets, the proband with two controls and the pre-enrollment case with the two non-proband controls.

Analyses of the DiMe data restricting to enrollment period cases and using all cases with probands eliminated as controls for non-enrollment period cases are given in the last two columns of Table 1. In both analyses, the estimated rate ratios for DR3 and DR4 are a little larger than the all siblings analysis, as would be predicted by the theory. Using non-enrollment period cases with proband elimination resulted in an additional 19 informative risk sets over restricting to enrollment period cases. In particular, in the latter analysis, only 15 pre-enrollment cases, out of the 498 (3%) total cases, had no controls. The two analysis methods were validated in extensive computer simulation studies described in the Technical Report. Each method yielded "unbiased" estimates of the rate ratio and the inverse information adequately estimated the variance of the estimates.

4 Discussion

In family studies such as the DiMe study, the restriction to enrollment period observation time as required by standard analysis methods results in a large loss of data and relaxation of this restriction is highly desirable. We found that the most significant way to increase the number of potentially informative risk sets is to allow any siblings who were known to be disease free at the age of the case to serve as controls. However, this departure from standard methods is at the cost of an additional exchangeability assumption, namely that the distribution of the covariates do not depend on calendar time or birth order. For genetic factors, this assumption seems quite reasonable but should be considered quite carefully when investigating environmentally or behaviorally determined factors. It obviously precludes any study of birth order or parental age effects.

We also found that the data set could be further augmented by including the risk sets formed by cases that occurred outside the enrollment period if the proband is excluded from serving as a control. Proband elimination is a well known ascertainment bias correction method in classical segregation analysis methods for single ascertainment. Our application of proband elimination has precedent in the conditional approach of Ewens and Shute (1986). In our context, conditioning on the "ascertainment event" results in dropping the proband from the pool of controls for non-enrollment period cases, since the proband could not have become diseased at that time. The main difference from the standard proband elimination method is that we only need to eliminate part of the proband's experience, that outside the enrollment period. In particular, we retain the most common case-control contributions, those in which the proband is the case. The difference between the two settings is that for rate ratio estimation, only the difference in covariate values between cases and controls must be representative of those in the population while for absolute risk estimation (needed in segregation analysis), the number of cases relative to the total subjects in the study must also be representative of the population.

Although we focused on fixed covariates and simple censoring because these were appropriate for the DiMe study, the analytic approach we used to study two-sibling families accommodates more complex situations. Specifically, the simple exchangeability is replaced by a “conditional” exchangeability assumption which essentially says that the distribution of covariates conditional on being under observation does not depend on birth order or calendar time. Further work is needed to assure that these are the only conditions necessary for general family structures.

Family studies of disease occurring at older ages or multigenerational studies could not be expected to be as complete as the DiMe study because of deaths due to other causes. Examination of the assumptions made in the theoretical development indicate that valid estimation is possible under quite general censoring so, in principle, the methods presented here apply. However, there is the critical issue of missing covariate information since it may be impossible to obtain blood or tissue samples for analysis of genetic factors in family members who have died. For these subjects, methods such as peeling or Gibbs sampling which average over the possible genotypes the subject could have had, given the known genotypes in the family, would be appropriate.

Since the grouped time analysis described at the end of Section 2 approximates the risk set approach, we expect that in situations where ignoring ascertainment results in little bias in rate ratio estimation, there will be little (additional) bias using the grouped time approach. However, parallel to our work, the contributions from families with non-enrollment period cases would be biased. While further work is needed in this area, the natural analog to our enrollment period cases approach is to form the family case-control set with enrollment period cases as cases and all other siblings, including the non-enrollment period cases, as controls. Our method which excludes the proband as a control for non-enrollment cases does not have an obvious parallel for the grouped time data. We note that the grouped time approach is no simpler computationally and introduces additional complications when there are time-dependent covariates due to the lack of a common "reference age." Further, the rare disease assumption is always assumed and, while this may be true in the entire population, the disease may not be rare for subgroups of families under study. For these reasons, we prefer the risk set approach.

The relatively small change in the rate ratio estimates in the DiMe analysis and in the simulation studies even when ascertainment is ignored altogether is a function of the simplex complete ascertainment method. More complex ascertainment, such as requiring two affected siblings, can result in severe bias if ignored. The bias under other ascertainment schemes and methods of bias correction are the topics of Ziogas (1996).

Acknowledgements

We thank Dr. Jaakko Tuomilehto of the National Public Health Institute of Finland for providing the data used for illustration and helpful comments on the manuscript. We also wish to acknowledge the Associate Editors for suggesting that we provide a more discursive, rather than mathematical, treatment of this problem. This work was supported by grant CA52862 of the United States National Cancer Institute.

Appendix: An analytic framework for investigating ascertainment bias

Let $N_j(t)$ be the usual counting process which indicates whether family member j has disease by age t and let $\mathcal{F}_{t-}$ be the disease, censoring, and covariate "history" for all family members up to but not including time t. With $dN_j(t) = 1$ denoting the event that subject j becomes diseased at time t, the intensity process associated with N_j is then, loosely, given by $\lambda_j(t)dt = E[dN_j(t)|\mathcal{F}_{t-}]$ which will generally depend on subject specific covariates. The time scale t is taken to be age, since this is the relevant time scale for comparing rates of disease in most chronic diseases, in particular, IDDM.

For family case-control studies ascertained through an affected proband, we do not observe N_j but rather $\tilde{N}_j$ which is N_j for members of ascertained families and is zero if the family is not ascertained. Let A be an indicator for the event that the family is ascertained into the study. The intensity process associated with $\tilde{N}_j$ is that of N_j conditional on A:

$$\begin{aligned} \tilde{\lambda}_j(t)dt &= \text{pr}[dN_j(t) = 1|\mathcal{F}_{t-}, A)] \\ &= A\frac{\text{pr}[dN_j(t) = 1, A = 1|\mathcal{F}_{t-}]}{\text{pr}[A = 1|\mathcal{F}_{t-}]} \\ &= A\frac{\text{pr}[A = 1|dN_j(t) = 1, \mathcal{F}_{t-}]\ \text{pr}[dN_j(t) = 1|\mathcal{F}_{t-}]}{\text{pr}[A = 1|\mathcal{F}_{t-}]} \\ &= \lambda_j(t)\ A\ \frac{\text{pr}[A = 1|dN_j(t) = 1, \mathcal{F}_{t-}]}{\text{pr}[A = 1|\mathcal{F}_{t-}]}dt \\ &= \lambda_j(t)\ A\ w_j(t)dt. \end{aligned} \tag{4}$$

Consistent with the definition of $\tilde{N}_j$, the inclusion of A in (4) assures that the intensity is zero for those families who are not ascertained. For ascertained families, the unconditional intensity $\lambda_j(t)$ is multiplied by a "distortion" due to ascertainment. Applying the standard martingale argument, the expectation of the score from (1) as a function of β is

$$E[U(\beta, t)] = \\ E\left\{\int_0^t \sum_{i=1}^n \left[Z_i(s) - \frac{\sum_j Y_j(s)Z_j(s)\exp(\beta Z_j(s))}{\sum_j Y_j(s)\exp(\beta Z_j(s))}\right] Y_i(s)\lambda_f(s)\exp(\beta_0 Z_i(s))\ A\ w_i(s)ds\right\} \tag{5}$$

where the expectation is over the covariates Z_j, censoring processes Y_j, and, in general, dates of birth and family size n. The derivation of an expression for the distortion under simplex complete ascertainment, and other analytic details, are given in the Technical Report.

References

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B*, **34**, 187–220.

Curtis, D. (1997). Use of siblings as controls in case-control association studies. *Annals of Human Genetics*, **61**, 319–333.

Eland-Johnson, R. (1971). *Probability models and statistical methods in genetics*. Wiley, New York.

Ewens, W. and Shute, N. (1986). A resolution of the ascertainment sampling problem: I. Theory. *Theoretical Population Biology*, **30**, 388–412.

Fisher, R. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, **6**, 13–25.

Langholz, B. and Huberman, M. (1999). Ascertainment bias in rate ratio estimation from case-sibling control studies of variable age-at-onset diseases: Supplementary material. Technical Report 108, Department of Preventive Medicine, Biostatistics Division, University of Southern California.

Langholz, B., Tuomilehto-Wolf, E., Thomas, D., Pitkäniemi, J., Tuomilehto, J., and The DiMe Study Group (1995). Variation in HLA-associated risks of childhood insulin dependent diabetes in the Finnish population: I. Allele effects at a, b, and dr loci. *Genetic Epidemiology*, **12**, 441–453.

Morton, N. (1959). Genetic tests under incomplete ascertainment. *American Journal of Human Genetics*, **11**, 1–16.

Self, S., Longton, G., Kopecky, K., and Liang, K.-Y. (1991). On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics*, **47**, 53–61.

Thompson, E. (1993). Sampling and ascertainment in genetic epidemiology: A tutorial review. Technical Report 243, Department of Statistics, University of Washington.

Tuomilehto, J., Lounanmaa, R., Tuomilehto-Wolf, E., Reunanen, A., Virtala, E., Kaprio, E., Åkerblom, H., and The DiMe Study Group (1992). Epidemiology of childhood diabetes mellitus in Finland - background of a nationwide study of Type 1 (insulin-dependent) diabetes mellitus. *Diabetologia*, **35**, 70–76.

Tuomilehto-Wolf, E., Tuomilehto, J., Cepaitis, Z., Lounanmaa, R., and The DIME Study Group (1989). New susceptibility haplotype for type 1 diabetes. *Lancet*, **2**, 299–302.

Ziogas, A. (1996). Correction for ascertainment bias in family studies. Doctoral dissertation, Department of Preventive Medicine, Biostatistics Division, University of Southern California.

Table 1: Characteristics of the analysis data sets and estimates of DR3 and DR4 rate ratios in a log-additive model (95% confidence intervals) for different approaches and restrictions to the risk sets.

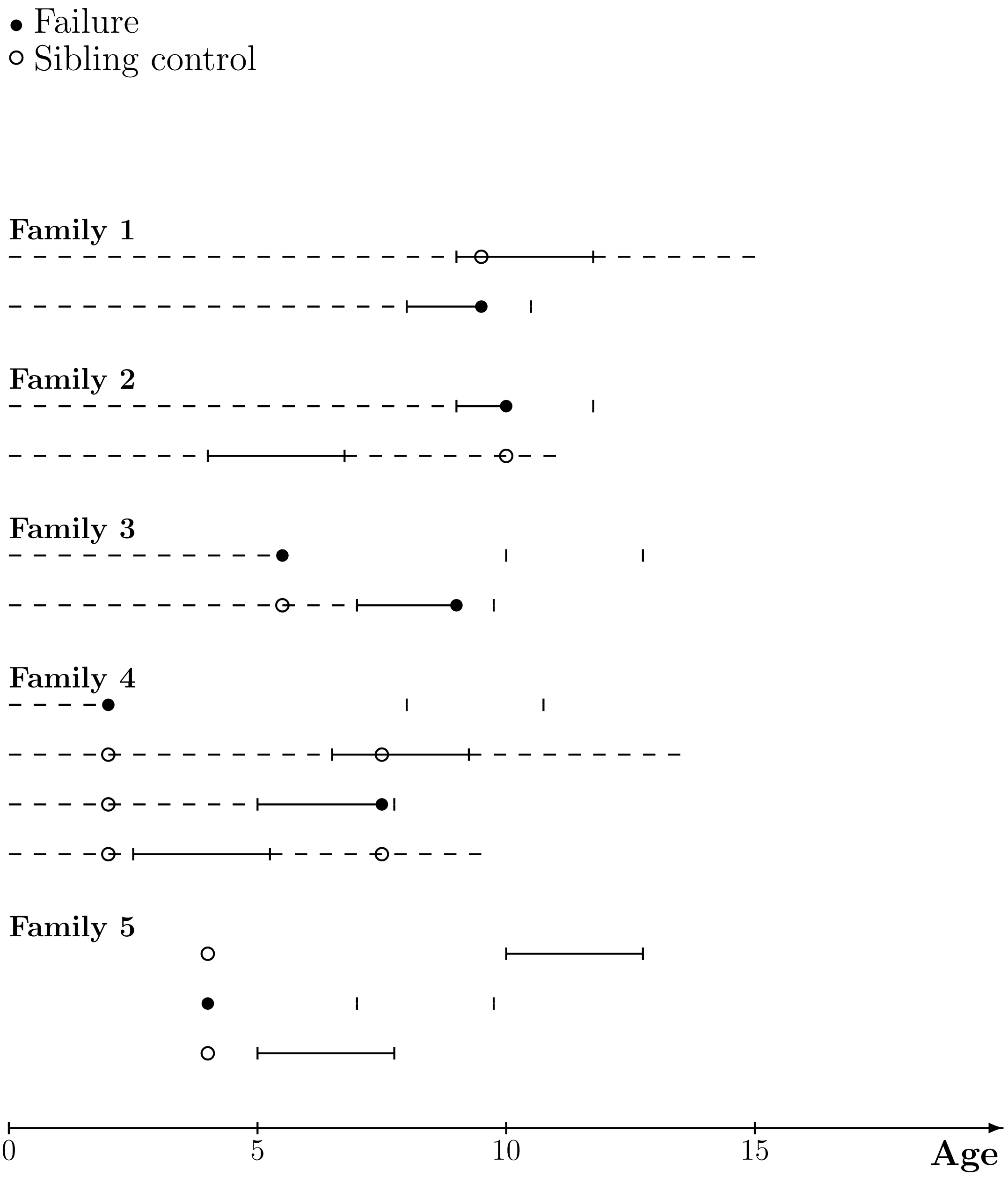
	All siblings	Censored observation period	Enrollment period cases	Proband excluded as a control
Rate ratios:				
DR3	4.6 (2.9-7.4)	16.0 (3.3-78.9)	5.0 (3.0-8.3)	4.9 (3.0-8.1)
DR4	8.1 (5.3-12.4)	7.5 (2.9-19.8)	8.2 (5.3-12.5)	8.4 (5.5-12.9)
Risk sets:				
Total informative[1]	498	114	464	483
Enrollment period case	464	114	464	464
Non-enrollment period case	34	0	0	19
Controls per risk set:				
1 control	345	110	307	326
2 controls	122	4	102	115
$>$ 2 controls	31	0	20	23

[1] Risk sets with at least one control.

Figure Legends

Figure 1. DiMe study as viewed on the age time scale.
Figure 2. Two-sibling family with non-overlapping enrollment periods: The four types of case-control pairs that can arise.
Figure 3. Two-sibling family with non-overlapping enrollment periods: The three additional types of case-control pairs that can arise when there is post-enrollment period follow-up.

Figure 1:



● Failure
○ Sibling control
Family 1
Family 2
Family 3
Family 4
Family 5
0
5
10
15
Age

Figure 2:

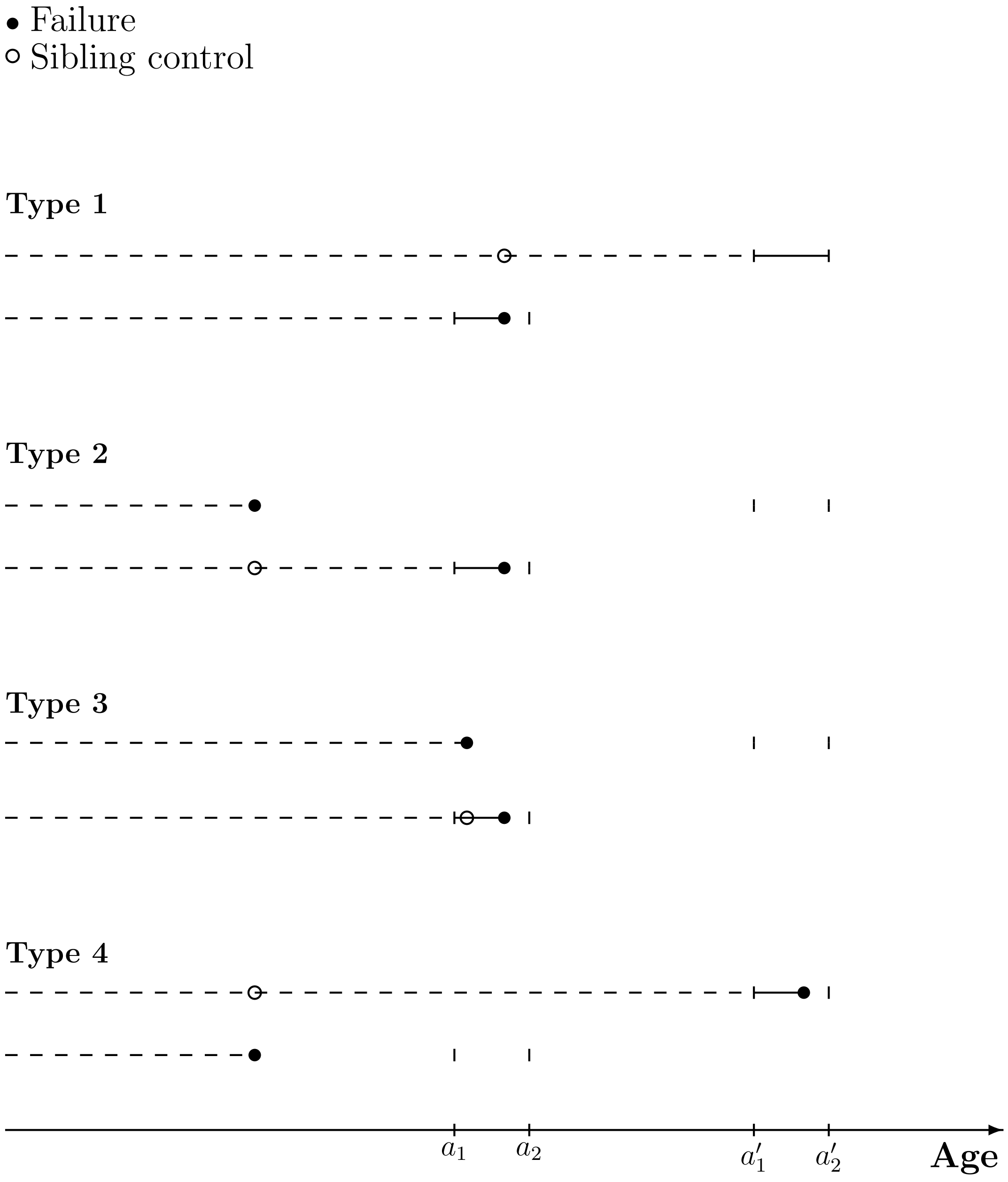
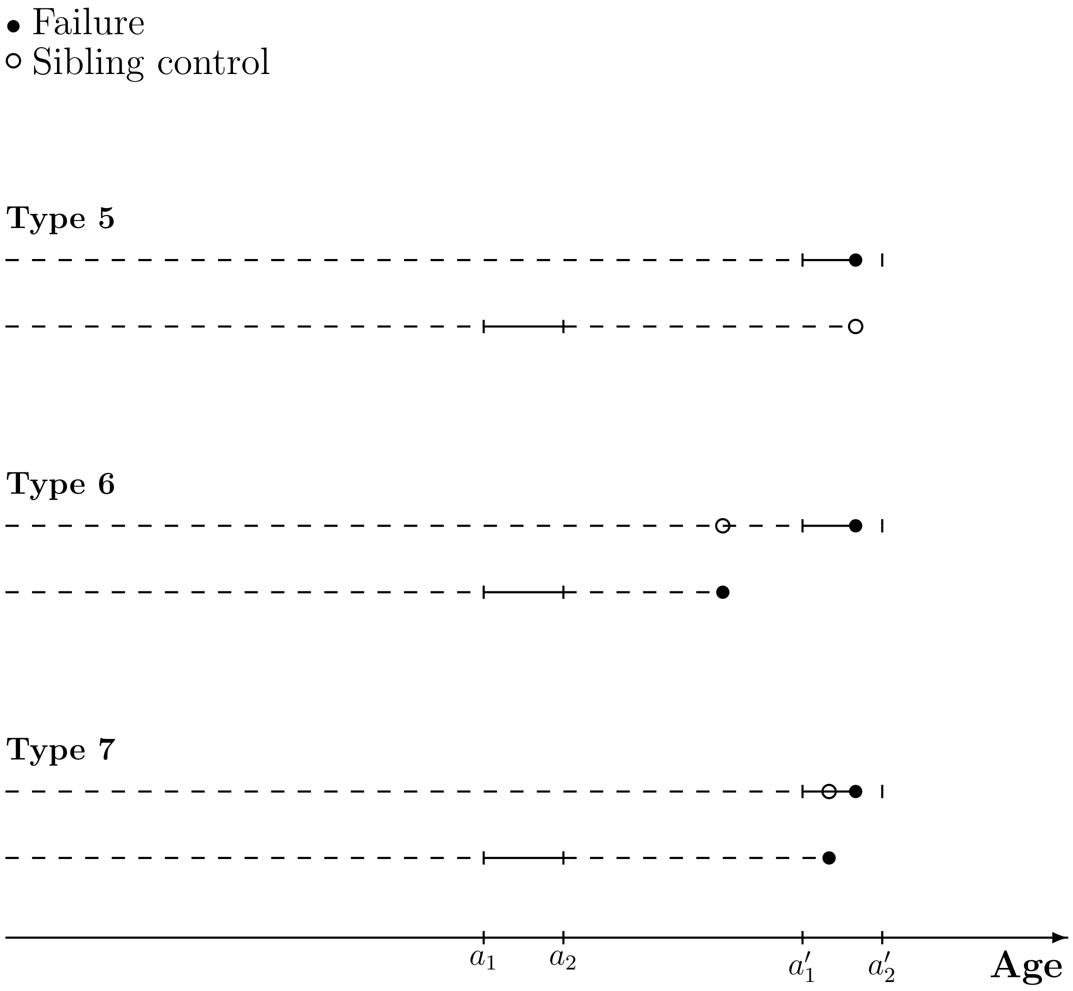

● Failure
○ Sibling control
Type 1
Type 2
Type 3
Type 4
a_1
a_2
a_1'
a_2'
Age

Figure 3:


● Failure
○ Sibling control
Type 5
Type 6
Type 7
a_1
a_2
a_1'
a_2'
Age