Published in final edited form as:

Stat Med. 2016 July 20; 35(16): 2815–2830. doi:10.1002/sim.6888.

Using Family Members to Augment Genetic Case-Control Studies of a Life-threatening Disease

Lu Chen[1,*], **Clarice R. Weinberg**[2], and **Jinbo Chen**[1]

[1]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104

[2]Biostatistics Branch, National Institute of Environmental Health, Research Triangle Park, NC 27709

Abstract

Survival bias is difficult to detect and adjust for in case-control genetic association studies but can invalidate findings when only surviving cases are studied and survival is associated with the genetic variants under study. Here, we propose a design where one genotypes genetically-informative family members (such as offspring, parents, and spouses) of deceased cases and incorporates that surrogate genetic information into a retrospective maximum likelihood analysis. We show that inclusion of genotype data from first-degree relatives permits unbiased estimation of genotype association parameters. We derive closed-form maximum likelihood estimates for association parameters under the widely used log-additive and dominant association models. Our proposed design not only permits a valid analysis but also enhances statistical power by augmenting the sample with indirectly studied individuals. Gene variants associated with poor prognosis can also be identified under this design. We provide simulation results to assess performance of the methods.

Keywords

Case-control study; Family genetic data; Non-ignorable missingness; Retrospective maximum likelihood; Survival bias

1. Introduction

In genetic association studies based on case-control data, many cases may not be available for genotyping only because they did not survive or they are too sick to participate. Such losses impose an ongoing challenge for research involving important debilitating or life-threatening conditions, because bias in inference, known as “survival bias” or healthy participant bias, will result if those losses are genetically informative. Survival bias, as one type of selection bias [1], has been recognized as a source of potential bias in many

*Lu Chen, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, 501 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, chenlu6@mail.med.upenn.edu, 1-215-350-8601.

Supplementary Materials

Web Appendices referenced in Sections 2 and 5 are available with this paper as the online material.

published genetic association studies (e.g., [2, 3, 4, 5]). It can also be a source of between-study heterogeneity in meta-analyses of genome-wide association studies [6], and may be responsible for failure to replicate some genetic association signals [7].

Current literature is sparse on study designs and statistical methods for addressing survival bias. Using simulation studies, Andersen et al. [7] evaluated the extent of bias by jointly modeling survival and association. The prospective study design was suggested as an unbiased alternative (e.g., [8]), but cohort studies are often infeasible. Alternatively, more stringent eligibility criteria can be applied to exclude some cases, e.g., based on recent diagnosis, (e.g., [9, 10]). But such constraints can lead to substantially reduced sample size and impose other kinds of selection that limit the generalizability of the study findings. Here, we develop a family-supplemented design (FSD) that permits correction and sensitivity analysis of survival bias in population-based case-control studies by genotyping family members. The central idea is that genotypes of family members of a deceased case can help provide information on the genotype of the deceased case, although the type of information depends on which family members are included. With our proposed strategy of genotyping cases and controls who are alive and also family members of representative deceased cases, we describe valid and efficient statistical methods for quantifying genetic associations both with disease and with survival given disease. We apply our method to a study of young-onset breast cancer where genotype data was available for slightly over half of the eligible case women by genotyping the available unaffected sister as a proxy for each unavailable case. We show that our method can greatly improve statistical power while simultaneously controlling healthy participant bias.

2. Method

We provide explanations for all notations in Section 6 Appendix Table 1 and Appendix Figure 1. We begin by considering only genetic effects. Let Y (coded as 1 or 0) denote case-control status. Suppose that among N_1 cases, n_1 have survived and provided genetic data. For simplicity of exposition, we assume that genetic data are available for one child, possibly together with that for the spouse (the other parent of that child), for each of the $m = N_1 - n_1$ deceased cases, or for a random subset of them. The method we describe below can be readily extended to incorporate data from multiple offspring or a mix of available family structures that could include a variety of genetically informative first-degree relatives, e.g. some siblings or parents of deceased cases, in lieu of offspring.

Assume that we have genetic data for N_0 controls, who are representative of the control population (i.e., those not studied are missing at random, and if stratified sampling was used the analysis will also be stratified). We wish to study the association between Y and a di-allelic autosomal single nucleotide polymorphism (SNP), with major/minor alleles A/a. Let G_i denote the SNP genotype data for subject i, and let G_i^f denote the genotype data of family members for the ith subject (only family members of deceased cases will be asked to provide genotype information). When one child (of the deceased case) and their spouse are accessible, $G_i^f = (G_i^c, G_i^p)$, where G_i^c and G_i^p denote their respective genotype data. When only one child or the spouse is accessible, $G_i^f = G_i^c$ or $G_i^f = G_i^p \cdot G_i^f$ is coded as missing when

no family member has been genotyped. We assume that a logistic regression model describes the association between Y and G:

$$p(Y=1|G)=\frac{e^{\beta_0+\beta f(G)}}{1+e^{\beta_0+\beta f(G)}}, \quad (1)$$

where $f(G)$ is a function of G that captures the penetrance model of interest, and β can be a vector. For example, for a co-dominant genetic model, $f(G)$ is a vector with two indicator functions $I_{(G=1)}$ and $I_{(G=2)}$, and $\beta = (\beta_1, \beta_2)$. For a log-additive model, $f(G)$ would be the number of copies (0, 1, or 2) of the minor allele a. Let S coded as 1/0 indicate the survival of a case (where $S = 1$ if survived). We assume that cases with genotype g have survival probability $q_g = p(S=1|G=g, Y=1)$. With a co-dominant model for survival, different probabilities, q_0, q_1 and q_2, are allowed, while q_1 and q_2 are equal when a dominant model is used.

When the genotype G is predictive of both risk of disease and survival given disease, one cannot assume that fitting model (1) with data from $n_1 < N_1$ cases and N_0 controls will yield unbiased estimates of the association parameter β. Under FSD, we consider that genotype data for one child, or the spouse (the child's biological parent), or both can be made available for all or a random subset of the m deceased cases. We develop a maximum likelihood estimation (MLE) method for estimating the log odds ratio parameter β with genotype data for the N_0 controls and n_1 cases and genotype data for the family member(s) of the m deceased cases. We shall assume Mendelian inheritance and that the influence of the spouse genotype on risk is only through the implicit genetic information that the spouse and child genotypes provide about the missing genotype of the case and not through any mechanism related to genetic population structure, or effects of the variant on fertility or assortative mating. The likelihood for the observed data can be written as:

$$\prod_{i=1}^{N_1}\left\{[p(S_i=1,G_i|Y_i=1)]^{S_i}\left[p(S_i=0,G_i^f|Y_i=1)\right]^{1-S_i}\right\}\prod_{i=N_1+1}^{N_1+N_0}p(G_i|Y_i=0). \quad (2)$$

Note that for now we are assuming for simplicity that only survival of cases and not that of controls is related to the genotype at the studied locus. However, this likelihood (the right-hand product factor) could easily be expanded to include an effect of the variant genotype on survival in those without the disease.

We further assume that the child's and spouse's genotypes at the locus under study are not related to the case's survival, conditional on the genotype of the case, i.e., $p(S_i=0|G_i,G_i^f,Y_i=1)=p(S_i=0|G_i,Y_i=1)$, and that the family members of cases whose genetic data are available are representative of case family members in the source population conditional on the case phenotype. Using a result from Satten and Kupper [11] to relate $p(G|Y=1)$ to $p(G|Y=0)$ for a rare outcome,

$$p(G|Y=1)=e^{\beta f(G)}\,p(G|Y=0)/\sum\nolimits_{G=g} e^{\beta f(g)}\,p(G=g|Y=0),$$

and by partitioning according to the composite events and applying the assumed models, the log of the observed data likelihood function (2) for a rare phenotype can be rewritten as:

$$\begin{aligned}\ell(\beta,q,p_a)=&\sum_{i=1}^{N_1}\left[S_i\log\left\{p\left(S_i=1|G_i,Y=1\right)e^{\beta f(G)}\,p\left(G_i|Y=0\right)\right\}\right]\\ &+\sum_{i=1}^{N_1}\left[(1-S_i)\log\sum_{G=g}\left\{p\left(S_i=0|G=g,Y=1\right)e^{\beta f(g)}\,p\left(G=g,G_i^f|Y=0\right)\right\}\right]\\ &-N_1\log\left\{\sum_{G=g}e^{\beta f(g)}\,p\left(G=g|Y=0\right)\right\}+\sum_{i=N_1+1}^{N_1+N_0}\log\left\{p\left(G_i|Y=0\right)\right\}.\end{aligned}\tag{3}$$

Under Hardy-Weinberg equilibrium (HWE), random mating, and Mendelian inheritance at the test locus in the population from which cases and controls arise, and assuming a rare phenotype, the probabilities $p(G, G^f|Y=0)$ and $p(G|G^f, Y=0)$ are functions of the MAF, p_a (Web Appendix 1). Let n_{10}, n_{11}, n_{12} denote the respective numbers of living cases (the second index) with genotypes *AA*, *Aa*, and *aa*. Let m_{jk} denote the number of deceased cases whose child's and spouse's genotype data, $G^c=j$ and $G^p=k$, is available, where j and k take values 0, 1, or 2 corresponding to genotypes *AA*, *Aa*, and *aa*, respectively. Note that not all combinations are possible; for example m_{02} and m_{20} cannot occur. Let m_{jc} denote the number of deceased cases whose child's genotype $G^c=j$ is available, while the spouse's genotype data is not.

Define inverse probability weighted cell counts $\hat{N_{10}} = n_{10}/\hat{q_0}$, $\hat{N_{11}} = n_{11}/\hat{q_1}$, and $\hat{N_{12}} + n_{12}/\hat{q_2}$, which can be seen as the predicted number of subjects with genotypes *AA*, *Aa*, and *aa* among the original total of N_1 case subjects. Note that, as one might hope, $\hat{N_{10}} + \hat{N_{11}} + \hat{N_{12}} + N_1$ (see Web Appendix). Under co-dominant coding for both association, $f(G)$, and survival (to study) conditional on disease, we obtain the maximum likelihood estimated odds ratio (OR) association parameters as

$$\hat{\beta}_1=\log\frac{\hat{N}_{11}\hat{N}_{00}}{\hat{N}_{10}\hat{N}_{01}}\text{ and }\hat{\beta}_2=\log\frac{\hat{N}_{12}\hat{N}_{00}}{\hat{N}_{10}\hat{N}_{02}},$$

where $\hat{N_{00}} = N_0(1-\hat{p_a})^2$, $\hat{N_{01}} = 2N_0\hat{p_a}(1-\hat{p_a})$, and $\hat{N}_{02}=N_0\hat{p}_a^2$ are the expected numbers of controls with genotypes *AA*, *Aa* and *aa*, respectively, under the rare disease approximation and HWE. These estimates are of the same form as those obtained from the two-by-three contingency table that cross-classifies the genotype and case-control status should the genotype be available for deceased cases, except that the "observed" counts are replaced by the "estimated" (for cases) or "expected" (for controls) counts. These estimators are also very similar to those obtained by Chen and Chatterjee [12] for estimating OR association parameters with case-control data under the constraint of HWE in the control population,

where they reported improved statistical efficiency when the observed numbers of controls were replaced by the "predicted" under HWE.

The expressions for $\hat{\beta_1}$ and $\hat{\beta_2}$ involve the estimated survival probabilities $\hat{q_0}$, $\hat{q_1}$ and $\hat{q_2}$. But it turns out that two of the score functions for the three suvival probabilities are equivalent, resulting in only two independent score equations. One cannot uniquely solve for three independent parameters from only two equations without additional constraints. Therefore, these three parameters as well as the two association parameters are not identifiable when a co-dominant model is adopted for both association and survival. Intuitively, this is because the genotype information of children and/or spouse is not fully informative in predicting cases' genotype. For example, when parents have genotype *Aa* and *aa*, one will expect that their child's genotype is *Aa* with probability 0.5 and *aa* with probability 0.5. However, when one child has genotype *aa* and the spouse has genotype *Aa*, one will only know that the case parent has genotype *Aa* or *aa* but cannot estimate the probabilities without taking advantage of the risk parameters for disease or survival.

When either the association model (1) or the survival model adopts an additional constraint, the parameters become identifiable. We note that children and spouses' genotype data are involved in formulas both for β and (q_0, q_1, q_2). Consequently, with genotype data from children and spouses, it is feasible to obtain maximum likelihood estimates that are consistent and at the same time assess the magnitude of survival bias. Below we provide estimates of OR parameters for association and survival probabilities under two commonly considered models. For each model, we consider three different scenarios of how family members of deceased cases are available: the general mixed scenario where genotype data for one child and spouse, or only for one child, or only for spouse is available for all or for a subset of deceased cases; one child and spouse scenario where genotype data for one child and spouse is available for some or all deceased cases; and one child only scenario where genotype data for only one child is available for all or some of the deceased cases. We also describe an extension of these results for studying gene-environment interactions. Here the estimates of log OR for association also resemble those if genotype data were available for cases, except that the genotype counts for cases and controls are replaced by the estimated (for cases) and the expected (for controls), respectively.

2.1 Modeling association as log-additive and survival as co-dominant

Under a log-additive association model for risk and a co-dominant model for survival, we obtained closed-form MLEs that incorporate family genotype data. The MLE for the log OR for association is

$$\hat{\beta}_1=\log\frac{(\hat{N}_{11}+2\hat{N}_{12})\hat{N}_{01}}{2(\hat{N}_{11}+2\hat{N}_{10})\hat{N}_{02}},$$

where again $\hat{N_{00}} = N_0(1-\hat{p_a})^2$, $\hat{N_{01}} = 2N_0\hat{p_a}(1-\hat{p_a})$, $\hat{N}_{02}=N_0\hat{p}_a^2$, and $\hat{N_{10}} + \hat{N_{11}} + \hat{N_{12}}$ are the inverse probability weighted counts, with $\hat{N_{10}} = n_{10}/\hat{q_0}$, $\hat{N_{11}} = n_{11}/\hat{q_1}$, and $\hat{N_{12}} + n_{12}/\hat{q_2}$ (Web Appendix). The MLEs for the three survival probabilities take the following forms:

$$\hat{q}_0=\frac{n_{10}}{N_1\,(1-\hat{p}_a^*)^2},\ \ \hat{q}_1=\frac{n_{11}}{2N_1\hat{p}_a^*\,(1-\hat{p}_a^*)},\ \text{and}\ \hat{q}_2=\frac{n_{12}}{N_1\,(\hat{p}_a^*)^2},$$

where $\hat{p}_a^*$, the estimated allele prevalence among affected individuals, takes the form $\hat{p}_a^*=(2M_a+n_{11}+2n_{12})/2N_1$. For the three scenarios to be described below, M_a and the estimated MAF $\hat{p_a}$ will take different forms and be defined differently. Note that M_a is not a parameter but an intermediate expression to represent half the estimated total number of copies of a in the ungenotyped cases, i.e. the part in $\hat{p}_a^*$ that differs among the three scenarios. Later in this section we will give an intuitive explanation for $\hat{p}_a^*$ and M_a. The scenarios considered in sections 2.1.2 and 2.1.3 are special cases of that to be described in section 2.1.1.

2.1.1 The General Mixed Scenario—We first consider a design where a representative sample of deceased cases has either a spouse or child or both available for genotyping. Let m_O denote the number of deceased cases who do not have child genotype data available. Let m_{1A} (m_{1a}) be the number of deceased cases for whom we have a genotyped child and who must have passed allele A(a) to the child, m_{1Aa} be the number of deceased cases for whom we have a genotyped child and the child has genotype Aa, m_{1*} be the number of deceased cases whose family members' genotypes provide no information in predicting cases' genotype. These counts are calculated as $m_{1A}=m_{00}+m_{01}+m_{12}+m_{0c}+m_{0+}+m_{12}$, $m_{1a}=m_{10}+m_{21}+m_{22}+m_{2c}=m_{10}+m_{2+}$, $m_{1Aa}+m_{1c}$, and $m_{1*}=m_{11}+m_O$. Then M_a mentioned above is the positive solution of the following quadratic equation, which is obtained directly by solving the likelihood score equation (See Web Appendix):

$$\frac{1-2p_a}{p_a}\,(m-m_{1*})\,M_a^2+m\left[m-m_{1*}-m_{_{1Aa}}-\frac{1-2p_a}{p_a}\,(m_{1a}+m_{_{1Aa}})\right]M_a-m^2m_{1a}=0.$$

There is no closed-form solution for the maximum likelihood estimator for the MAF p_a. Therefore we obtained the profile likelihood for p_a based on the closed forms of $\hat{q_0}$, $\hat{q_1}$, $\hat{q_2}$ and $\hat{\beta_1}$, and maximized it numerically. Because of the correlation between parental genotypes conditional on the child's genotype, incorporating the spouse genotype data should increase the precision of estimating both the association parameter β_1 (through more precise estimation of p_a) and the survival probabilities. We note that the scenarios considered in Sections 2.1.2 and 2.1.3 are both special cases of this general scenario. When for every deceased case only one child is recruited, $m_{jk}=0$ for all j and k. When for every deceased case both the spouse and one child are recruited, $m_{jc}=0$ for all j.

2.1.2 The One Child and the Spouse Scenario—Under this scenario, we assume that genotype data for one child and the spouse are available for all or a representative sample of the deceased cases. We calculated M_a and derived the MAF estimate $\hat{p_a}$ as:

$$M_a=\frac{mm_{1a}}{m-m_{1*}}=\frac{m(m_{2+}+m_{10})}{m-m_{1*}},\qquad \hat{p}_a=\frac{N_{01}+2N_{02}+m_{1p}+2m_{2p}}{2(N_0+m_{+p})}.$$

Here $m_{1a} = m_{10} + m_{2+}$ according to section 2.1.1, m_{1p} and m_{2p} are the numbers of deceased cases who have spouse genotype *Aa* and *aa*, respectively, and $m_{+p} = m_{0p} + m_{1p} + m_{2p}$. Note that the MAF estimate $\hat{p_a}$ uses data for both controls and spouses of the deceased cases. This is because the spouses do not have the disease and can be treated as an independent source of data for estimating the MAF.

The terms M_a and $\hat{p}_a^*$ have intuitive interpretations. For the m_{2+} deceased cases whose child's genotype is *aa* and m_{10} cases whose child has genotype *Aa* and spouse has genotype *AA*, each case must have passed an "*a*" allele to the child. Similarly, for the m_{0+} deceased cases whose child's genotype is *AA* and the m_{12} cases whose child has genotype *Aa* and spouse has genotype *aa*, each case must have passed an "*A*" allele to the child. Thus there are $m_{2+} + m_{10}$ and $m_{0+} + m_{12}$ deceased cases who passed "*a*" and "*A*" to the offspring, respectively. Note that $m - m_{1*} = m_{2+} + m_{10} + m_{0+} + m_{12}$. Since "*A*" and "*a*" have equal probability to be transmitted from a heterozygous parent to his/her child, the total number of deceased cases who have passed "*a*" alleles can be estimated as $m\frac{m_{2+}+m_{10}}{m-m_{1*}}$, which is M_a. The number of surviving cases who have passed "*a*" alleles to the offspring is estimated as $\frac{n_{11}}{2}+n_{12}$. Then the estimated total number of cases passing "*a*" alleles to their offspring is $M_a+\frac{n_{11}}{2}+n_{12}$. Define the estimated MAF among cases as $\hat{p}_a^*=\hat{p}(a|Y=1)$ then

$$\hat{p}_a^*=\frac{M_a+\frac{n_{11}}{2}+n_{12}}{N_1}=\frac{2M_a+n_{11}+2n_{12}}{2N_1}.$$

Note that $\hat{p}_a^*$ is a consistent estimator for the MAF in the case population also because the genotype in the case population conforms to HWE under the log-additive risk model.

2.1.3 One Child only Scenario—We next consider a design where a representative sample of cases has a child available for genotyping and spouses are not genotyped. Sometimes it may be feasible to obtain genotype data only for offspring, but not spouses. Let m_{0c}, m_{1c}, m_{2c} be the respective number of children with genotype *AA*, *Aa*, *aa,* and $m_O = m - m_{+c}$ the number of deceased cases without child genotype data. In this case, M_a is the positive solution of the following quadratic equation:

$$\frac{1-2p_a}{p_a}(m-m_O)\ M_a^2+m\left[m_{0c}+m_{2c}-\frac{1-2p_a}{p_a}(m_{2c}+m_{1c})\right]\ M_a-m^2m_{2c}=0.$$

No closed form estimate exists for the MAF. Therefore we obtained the profile likelihood for p_a, which is a function of the closed-form estimates $\hat{q_0}$, $\hat{q_1}$, $\hat{q_2}$, and $\hat{\beta_1}$, and maximized it numerically as for the general scenario.

2.2 Use of dominant coding for both association and survival

We next consider the use of a dominant model for survival. This version may be needed in situations, e.g., if the frequency of the minor allele under study is low, where a co-dominant

model for survival becomes impractical due to sparse data and more parsimonious models become necessary. With a dominant model for both association and survival, the MLEs are calculated using methods similar to those of section 2.1 as follows. For all the three scenarios considered in section 2.1, the estimated OR association parameter takes the form

$$\hat{\beta}_1=\log\frac{\left(\hat{N}_{11}+\hat{N}_{12}\right)\,\hat{N}_{00}}{\left(\hat{N}_{01}+\hat{N}_{02}\right)\,\hat{N}_{10}},$$

where $\hat{N_{00}}$, $\hat{N_{01}}$, and $\hat{N_{02}}$ are the predicted genotype counts in controls under HWE as described previously,

$$\hat{N}_{10}=N_1\,(1-\hat{p}_a^{**})\,,\quad \hat{N}_{11}=N_1\hat{p}_a^{**}\frac{n_{11}}{n_{11}+n_{12}},\quad \hat{N}_{12}=N_1\hat{p}_a^{**}\frac{n_{12}}{n_{11}+n_{12}}.$$

Here $\hat{p}_a^{**}=\frac{(2-\hat{p}_a)\;M_a+n_{11}+n_{12}}{N_1}$, and M_a takes the same forms for the three scenarios as in section 2.1.1, 2.1.2 and 2.1.3, respectively. Note that M_a has the same interpretation as in section 2.1.2. Then $(2-\hat{p_a})M_a$ can be interpreted as the estimated number of deceased cases who have at least one copy of the minor allele. Since n_{11} + n_{12} is number of surviving cases who carry the minor allele, $\hat{p}_a^{**}$ is the estimated proportion of cases who carry the minor allele. Under this dominant model, we obtained $\hat{p_a}$ based on its profile likelihood.

2.3 Extension of the Above Results for Assessing G-E interactions

The estimation and testing of multiplicative *G-E* interaction effects will also be subject to survival bias. Suppose that a binary exposure *E* influences the occurrence of disease and might also influence survival conditional on disease. We assume E is available for all N_0 +N_1 study participants, regardless of availability of genotype data. If survival is multiplicative in *E* effects and *G* effects, and if both survival and association depend on the same binary coding for *G* or the association model (but not the survival model) depends on co-dominant coding for *G*, then the naïve estimate of the multiplicative *G-E* interaction OR parameter from standard logistic regression analysis and the accompanying standard error estimate are valid according to theoretical results for two-phase studies [13]. Of course the naïve main-effect OR estimates for *G* and for *E* would typically be biased. On the other hand, if survival depending on *G* follows a co-dominant model and the occurrence of disease depends on *G* through a log-additive model, then naïve estimates of both the main and interaction effect OR parameters would often be biased. The methods in the first two subsections can be extended to obtain consistent estimates of all OR parameters. Consider a simple logistic regression model for disease occurrence that quantifies the joint effect of a di-allelic SNP and a binary environmental variable:

$$p(Y=1|G,E)=\frac{e^{\beta_0+\beta_g f(G)+\beta_e E+\beta_I f(G)\times E}}{1+e^{\beta_0+\beta_g f(G)+\beta_e E+\beta_I f(G)\times E}},$$

where $f(G)$ is defined as in Equation (1), and E takes the value 1 or 0. Let N_{ijk} denote the total number of subjects with $Y=i$ ($i=0, 1$), $G=j$ ($j=0, 1, 2$), and $E=k$ ($k=0, 1$). Let $\hat{N}_{1jk}$ be the "predicted number" of cases defined as above through inverse probability weighting. Under the assumptions of G/E independence and HWE in the source population, and initially assuming that survival depends only on G and not on E, we derived estimates of all association parameters. Note, however, that G/E independence is only assumed here for convenience of demonstrating our method and not needed for our general proposed design. If $f(G)$ expresses log-additive coding and $h(G)$ co-dominant coding, we obtain the same estimates as those in Section 2.1 for the population MAF p_a and the survival probabilities (q_0, q_1, q_2) and

$$e^{\hat{\beta}_e}=\frac{\hat{N}_{1+0}N_{0+0}\left(\hat{N}_{111}+2\hat{N}_{101}\right)^2}{\hat{N}_{1+1}N_{0+1}\left(\hat{N}_{110}+2\hat{N}_{100}\right)^2},\ e^{\hat{\beta}_g}=\frac{(1-\hat{p}_a)\left(\hat{N}_{110}+2\hat{N}_{120}\right)}{\hat{p}_a\left(\hat{N}_{110}+2\hat{N}_{100}\right)},\ e^{\hat{\beta}_I}=\frac{\left(\hat{N}_{110}+2\hat{N}_{100}\right)\left(\hat{N}_{111}+2\hat{N}_{121}\right)}{\left(\hat{N}_{111}+2\hat{N}_{101}\right)\left(\hat{N}_{110}+2\hat{N}_{120}\right)}.$$

Under a dominant model for both $f(G)$ and $h(G)$, estimates of $\hat{q}_0$, $\hat{q}_1$, $\hat{q}_2$ are the same as those in Section 2.2, and the estimates of the OR parameters are

$$e^{\hat{\beta}_e}=\frac{N_{0+0}\hat{N}_{101}}{N_{0+1}\hat{N}_{100}},\ e^{\hat{\beta}_g}=\frac{(1-\hat{p}_a)^2(\hat{N}_{110}+\hat{N}_{120})}{\hat{p}_a(2-\hat{p}_a)\hat{N}_{110}},\ \text{and}\ e^{\hat{\beta}_I}=\frac{\hat{N}_{100}(\hat{N}_{111}+\hat{N}_{121})}{\hat{N}_{101}(\hat{N}_{110}+\hat{N}_{120})}.$$

The two interaction parameter estimates above only involve data for cases and their families and reduce to case-only estimates [14] under G/E independence in the absence of survival bias. The association parameter estimators resemble those derived in Chen et al. [15] in the absence of survival bias, except that the observed counts in cases are replaced by the estimated counts obtained through inverse-probability weighting. Similar results can be obtained when survival depends on E both through its main effect and interaction with $h(G)$, where survival probabilities at each G/E combination need to be estimated.

3. Simulations

We performed numerical studies to evaluate the performance of the proposed methods in realistic scenarios for assessing marginal effects of G. Imposing our rare phenotype assumption, we generated SNP genotypes for controls under HWE, and for cases from the derived distribution

$$p\,(G|Y=1)=\frac{e^{\beta f(G)}\,p\,(G|Y=0)}{\sum_{g=0,1,2}e^{\beta f(g)}\,p\,(G=g|Y=0)},$$

where the coding we impose is either log-additive, $h(G)$= the number of copies of the minor allele, or dominant, $h(G)$=I(number of copies of the minor allele > 0). We generated the indicator for failure to survive, S=0, from the mixture distribution

$$p(S=0|G,Y=1)=\rho+(1-\rho)\frac{e^{h(G)}}{1+e^{h(G)}},\ 0<\rho<1.$$

Note that we do not fit this mixture distribution directly but make use here of the fact that without knowing the true missing mechanism, we are still able to obtain unbiased estimates of the survival probabilities and disease odds ratios as long as probabilities of survival are allowed to be different for people with different genotypes. We used a log-additive $h(G) = \alpha_0 + \alpha_1 G$ when generating S but fitted the less restrictive co-dominant model for $h(G)$ when the association model was log-additive, and used dominant coding for both fitting and generating S when the true association model was dominant. We refer to $e^{\alpha 1}$ as the "OR for death" below. Thus the true $p(S=0 \mid G,Y=1)$ depends on G in a co-dominant or dominant fashion. Cases with $S = 0$ were treated as "deceased" in the analysis. We used random survival as the gold standard for comparison, which was generated by setting $\alpha_1 = 0$. Then we generated two separate indicator variables R_c and R_p for the general mixed scenario from the models $p(R_c = 1| S = 0,G,Y = 1) = \gamma_c$ and $p(R_p = 1| S=0,G,Y=1) = \gamma_p$, where $R_c = 1$ and $R_p = 1$ indicated that the genotype data was available for the child and spouse. Similarly, for scenarios where the child/spouse scenario genotype data for the child and spouse are available or not, as a pair, we generated only one indicator R from $p(R = 1| S = 0,G,Y = 1) = \gamma$. For the child-only scenario, we generated indicator R_c from $p(R_c = 1| S = 0,G,Y = 1) = \gamma_c$. Below we refer to γ_c, γ_p, and γ as "recruiting probabilities". We used a MAF of 0.2 and for α_0 a value of −2.0, and each selected scenario was simulated 5,000 times.

Table 1 presents results showing the magnitude of survival bias for an analysis that simply excludes data for deceased cases, based on 5000 simulations. When the tested variant was independent of or only modestly associated with death ($e^{\alpha 1} < 1.4$), the bias in estimation was small, and the type I error rates were close to the nominal 0.05 level. When the variant was more strongly associated with death, the type I error rate became notably inflated. When only 30% of the missing genotype data were missing due to reasons not related to G and 70% were missing due to G-related causes, under the log-additive model for association and co-dominant model for survival, with the OR for death being 2.0, the missing genotype data rates corresponding to the three genotypes AA, Aa, and aa were 0.38, 0.45, and 0.55 respectively. The observed type I error rate corresponding to a nominal 0.05 was 0.145 when deceased cases were naively ignored in standard likelihood ratio tests of association. Our method maintained nominal type-I error rates in all scenarios. In general, the bias in the naïve estimation of the association OR parameter was small. For example, the mean estimate was 1.58 when the true value for association was 1.8 and the OR for death ($e^{\alpha 1}$) was 2.0 (data not shown). Similar observations were made when a dominant model was adopted for both survival and association (with the data simulation and the analytic models corresponding). Table 2 presents the average log OR estimates and the estimated standard deviations. The average estimates obtained by our method were close to the true value. As

expected, the child and spouse scenario resulted in the lowest standard deviations, while the child-only scenario had the highest standard deviations. The estimates without accounting for survival (rows “Naïve”) appeared to be biased, but the bias was small.

Figure 1 displays the powers under different models, scenarios and recruiting probabilities. In both the general mixed (“Mixed”) and child and spouse (“Child+Spouse”) scenarios, the recruiting probabilities for children and spouses were assumed to be 1, 0.7 or 0.5. In the child-only (“Child”) scenario, children had recruiting probabilities 1, 0.7 or 0.5. Except for the three scenarios in sections 2.1, we considered two additional scenarios: the ideal one where all cases survived, and the general mixed scenario where the recruiting probability of spouses was half of that of children (“Mixed_less Spouse”). That is, when the recruiting probabilities γ_c for children were assumed to be 1, 0.7 or 0.5, the corresponding recruiting probabilities for spouses γ_p were 0.5, 0.35, and 0.25, respectively. The power appeared to be in increasing order from Child-Only, Mixed_less Spouse, Mixed, Child + Spouse to Ideal. When the recruiting probability was 1 in the log-additive model, the powers of both the Mixed and Child + Spouse scenarios were close to the ideal. Not surprisingly, the power of all scenarios decreased with decreasing recruiting probabilities. For example, under the model of log-additive coding for association (odds ratio 1.4) and co-dominant coding for survival, in the Mixed scenario, the powers corresponding to recruiting probabilities 1, 0.7 and 0.5 were 0.848, 0.723 and 0.606, respectively.

Interestingly, although the spouse genotype itself provides no information in predicting the case’s genotype, inclusion of spouses together with the child genotype can improve the power noticeably. Consider testing the association of a SNP under the log-additive coding with OR 1.4. Consider that genotype data is available for 500 controls and that the survival probabilities are 0.62, 0.59 and 0.55 for cases with genotype AA, Aa, and aa respectively. Assume that genotype data are also available for family members of 70% of deceased cases. If 70% of children of the deceased cases provided genotype information, the power was 0.565. If 70% of spouses of the deceased cases were also randomly available, then on average, about half (49%) of the deceased cases have both children and spouses genotype. The study power under this setting was 0.723. If genotype data were jointly available for the child and spouse as a pair for 70% of the deceased cases, the study power increased to 0.782. A natural question that arises under a fixed budget is: which one, recruiting x children or $x/2$ Child + Spouse pairs, would yield higher study power? As stated previously, the power was 0.565 with genotype data for 70% children. The power increased to 0.595 with genotype data for 35% child and spouse pairs. Similarly, compared to the power 0.640 when genotyping only children of all the deceased cases, genotyping both the child and the spouse of half of the deceased cases yielded higher power, 0.682.

Even when survival was independent of G but the possible dependence of survival on G was nevertheless modeled, our design achieved an increase in power by incorporating family genotype data and exploiting the HWE assumption, partly through the improved estimation of the MAF. With $\beta_1 = \ln(1.2)$ and $\beta_1 = \ln(1.4)$ (data not shown), the power by the likelihood ratio test based on standard logistic regression with surviving cases only was 0.322 and 0.792, respectively, which are lower than those with the proposed method, 0.356 and 0.853.

When both models were dominant, the corresponding powers were 0.247 and 0.658 versus 0.257 and 0.676, respectively.

4. Application to a Study of Breast Cancer Genetics

We analyzed a dataset that was derived from the Two Sister Study (http://sisterstudy.niehs.nih.gov/English/studies.htm). Although the study design was based on young-onset breast cancer in the daughters, the analysis presented here is based on the occurrence of breast cancer in the mothers. The dataset included 291 women who had in the past been diagnosed with breast cancer and 645 female controls who had never been diagnosed with breast cancer. Genetic information on 6 SNPs from the gene FGFR2 (rs3750817, exm.rs2981579, rs2981578, exm.rs2981575, rs1219648, rs2981582) were available for all 645 controls but for only 126 cases. The additional 165 cases who could not be genotyped had a daughter (or daughters) or spouse who had been genotyped as part of the study. Therefore, for those 165 mothers we used the daughters' genotype data to supplement their missing mother's genotype, using one from each family (preferring an unaffected daughter if two were available). Among the 165 missing cases, 29 had both a daughter and the daughter's father available for genotyping and 135 had only a daughter available. The remaining missing case had no family members available and was omitted from analysis. For each SNP, we conducted two analyses, each applying log-additive coding for risk: (1) we fit a logistic regression model, simply omitting the ungenotyped cases (column "Naïve"); and (2) we applied the proposed FSD (assuming co-dominant coding for missingness) to the dataset where we made use of the proxy family members for cases with missing genotypes, enabling inclusion of 290 cases instead of 126 cases (column "FSD"). Note that although the mechanism for missing genotype was, for simplicity, assumed to be death in the description of the methods given above, we here think of it more broadly as genotype-related missingness.

A test of Hardy-Weinberg equilibrium based on controls yielded reassuringly hefty p values for all 6 of the SNPs (data not shown), verifying an assumption needed for FSD. Results based on the three analytic methods applied for assessing effects of FGFR2 on risk of breast cancer are shown in Table 3. We calculated the estimates for the OR parameters and log OR parameters with standard errors for each SNP with both methods, and calculated tail probabilities for the Z statistics as the log OR divided by the standard error (column "P-value"). In Table 4, we provide the "survival" probabilities (more accurately, recruitability) for each of the three genotypes for each SNP.

Although the standard errors of the Naive method are slightly smaller than those from FSD, with this example the FSD method always results in more significant P-values. For example, for the first SNP rs3750817, the Naive method indicated no association between breast cancer and this SNP, while the FSD method obtained a marginally significant P-value. Similarly, for the last SNP rs2981582, the P-value from FSD (0.003) is much smaller than that from the Naive method (0.028), where the latter is not significant after the Bonferroni adjustment of testing 6 SNPs.

The FSD method also enables us to explore whether missingness is related to genotype. Results are shown in Table 4, where "a" denotes the minor allele (the allele with prevalence less than 0.5). All of the p values for testing equality of the recruitability probabilities were above 0.24. Therefore, missingness did not seem to depend on the genotype for these FGFR2 SNPs. We then conducted a stage 2 analysis where we constrained $q_0 = q_1 = q_2$. Results are shown in Table 3, columns "FSD^f". Here we have sacrificed protection against survival bias, but retained our ability to capture the supplemental information derived from proxy family members. Note that, relieved of its need to estimate those three parameters, the analysis now returned smaller standard errors than were achieved by the original naïve analysis that was shown in Table 3. It is interesting to observe the estimates by "Naïve" and "FSD^f" always yielded similar estimates, which may differ noticeably from that by FSD. This difference was due to the differential "survival" probabilities for women carrying different genotypes for each SNP, as shown in the "Survival Probabilities" column in Table 4. Therefore, we prefer to interpret association findings based on FSD, instead of "FSD^f".

5. Discussion

The proposed design provides a useful approach for assessing and correcting survival bias in population-based case-control genetic association studies and improving power by enrolling offspring and spouses of deceased cases and making proxy use of their genotypes. The approach performed well for assessing both marginal genetic effects and gene-environment (*G*-*E*) interactions (unreported simulation results). It also yielded unbiased estimates of association parameters for survival (data not shown). In addition to HWE, we make the usual assumptions that cases and controls are drawn from the same source population and that the disease is rare. The current exposition considers a single di-allelic SNP in relation to association and survival and the same study design can be extended to study multiple genetic and environmental factors, through application of the EM algorithm. Our approach introduces a useful solution to a thorny non-ignorable missingness problem. Because it is based on a retrospective likelihood, mis-specification of the nuisance survival model may lead to biased estimates and inflated type-I error rate (Web Appendix Table 2). Therefore, a co-dominant model, which is always a correct model, should be fitted for survival whenever the data allows. We fit the log-additive or dominant model for survival only when the data is too sparse, i.e, the number of subjects for a genotype category is too small to fit a co-dominant model.

The family-supplemented design does have limitations. Because of identifiability issues one cannot use co-dominant models for both association and survival. The identifiability issue can presumably be resolved if genotype data could be made available for a random subset of deceased cases (not likely in practice), but the issue cannot be addressed by genotyping family members or obtaining richer marker genotype data. It is desirable to consider other design options to address this challenge. In addition, extending our proposed method to incorporate the survival time of the deceased cases may also prove helpful.

As is true generally for case-control genetic analyses, our method requires that there be no population stratification. We also implicitly assume there is no assortative mating related to the SNP under study and that the SNP is not related to fertility (i.e. the availability of a child

to serve as a genetic proxy). However, we expect that a more robust approach can be devised for settings where there may be population stratification, using genetic control for population stratification, if we can approximate that the population that cases and controls are sampled from contains a finite number of discrete strata (e.g., [17]), mating is restricted to within-strata unions, and HWE holds within each stratum. Ancestry informative markers would need to be obtained for controls, live cases, and family members of deceased cases. Joint inference would be needed for association parameters, survival probabilities, stratum membership, and minor allele frequency (MAF) within each stratum. This task, at least computationally, is nontrivial and worth investigating in future research. Our proposed design can potentially be generalized to address survival bias in family-based case-control studies, which are largely free of bias from population stratification.

We focused the theory of our method on the scenario where genotype data can be obtained for one child and/or the spouse for each deceased case in a representative subsample. But in many realistic settings some deceased cases may not have a child available for genotyping but might have one or both parents available, while still others might provide a sibling, a child, or maybe a child and one parent. Under HWE and standard Mendelian transmission, the Expectation-Maximization (EM) algorithm [16] can be used to handle missing case genotypes under diverse family structures. In particular, the method is readily extensible to include multiple children with or without the spouse. If the number of family members recruited for each deceased case is specified *a priori*, then a relevant question is which family members are the most informative if they are equally recruitable. Clearly the spouse alone would not be very informative, whereas a single offspring alone would provide somewhat limited but informative genotype data as shown in the table and in the numerical studies. Other choices are not so clear. Consider two children versus one child and the spouse. When the two children have genotypes AA and aa, one can infer that the genotype for both parents is Aa. But other combinations of the genotypes may not be as informative as the same combination for one child and spouse (for example, if the two children have the same genotype). In some circumstances *a priori* knowledge of the biology might be needed. For example, if the gene under study may be related to longevity, the availability of the spouse may itself be selective (genotype-dependent), and children or siblings might be preferred for that reason. Nevertheless, we explored the most plausible scenarios in our theoretical and numerical studies, as recruitment of multiple children for a case may be practically challenging even if it were more efficient.

For conditions with young age at onset, such as birth defects, children and spouses will not be available at all and one would need to use the parents or siblings of cases as genetic informants. The genotypes of deceased (or therapeutically aborted) cases then become missing data and amenable to likelihood methods for estimation and testing through the EM algorithm.

We have focused on genotype data for a single SNP at a time. But genotype data at nearby markers could be helpful for infering the unobserved genotype data at that SNP and could usefully be incorporated into the likelihood analysis [18]. Consider another SNP with alleles B and b that is so close to our SNP of interest (which has alleles A and a) that the probability of recombination from parent to child (i.e. in a single meiosis) can be taken to be 0. Suppose

the case had a child with genotypes *Aa*/*Bb* and spouse *Aa*/*BB* at the two loci. The child must carry either haplotypes *AB* and *ab* or haplotypes *Ab* and *aB*, while the spouse carries haplotypes *AB* and *aB*. The child must have inherited either *AB* or *aB* from the spouse. If *Ab* does not exist in the population, the case must have a copy of haplotype *ab*. Without genotype data for the second locus, the child and spouse genotype at the first locus, *Aa*/*Aa*, are less informative for the missing parent case's genotype. Presumably, again the EM algorithm could be used to take advantage in this way of the LD structure in the genome for imputing missing genotypes, though the details are beyond our present scope.

In conclusion, when studying the role of genetics in causing a disease that can be fatal, survival bias can result in nonignorable missingness of genotype information. Under plausible assumptions, the genotyping of family members of deceased cases can both correct that bias and improve power for detecting genetic associations. Our methods are simple to implement and therefore can be applied to analyze genomewide association studies. Software for implementing this method is available from the authors upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. David Umbach for useful comments. This research was supported in part by the NIH grants ES016626/ES020811 (for LC and JC) and Intramural Research Program of the NIH, National Institute of Environmental Health Sciences, under project number Z01 ES040007 (for CRW) and by Susan G. Komen for the Cure (grant FAS 0703856).

References

1. Rothman, KJ.; Greenland, S.; Lash, TL. Modern Epidemiology. 3. Lippincott-Williams-Wilkins Publishers; Philadelphia, PA: 2008.
2. Chiu CJ, Conley YP, Gorin MB, Gensler G, Lai CQ, Shang F, Taylor A. Associations between Genetic Polymorphisms of Insulin-like Growth Factor Axis Genes and Risk for Age-Related Macular Degeneration. Investigative Ophthalmology & Visual Science. 2011; 52:9099–107. [PubMed: 22058336]
3. Lazarevic G, Milojkovic M, Tasic I, Najman S, Antic S, Stefanovic V. PC-1 (ENPP1) K121Q polymorphism in overweight and obese type 2 diabetic coronary heart disease patients. Acta Cardiologica. 2008; 63:323–30. [PubMed: 18664022]
4. Lacour RA, Westin SN, Meyer LA, Wingo SN, Schorge JO, Brooks R, Mutch D, Molina A, Sutphen R, Barnes M, Elder J, Teoh D, Powell CB, Choubey V, Blank S, Macdonald HR, Brady MF, Urbauer DL, Bodurka D, Gershenson DM, Lu KH. Improved survival in non-Ashkenazi Jewish ovarian cancer patients with BRCA1 and BRCA2 gene mutations. Gynecologic Oncology. 2011; 121(2):358–63. [PubMed: 21276604]
5. Williams P, Pendyala L, Superko. Survival bias and drug interaction can attenuate cross-sectional case-control comparisons of genes with health outcomes. An example of the kinesin-like protein 6 (KIF6) Trp719Arg polymorphism and coronary heart disease. BMC Med Genet. 2011; :12–42. DOI: 10.1186/1471-2350-12-42 [PubMed: 21247423]
6. Gogele M, Minelli C, Thakkinstian A, Yurkiewich A, Pattaro C, Pramstaller PP, Little J, Attia J, Thompson JR. Methods for Meta-Analyses of Genome-wide Association Studies: Critical Assessment of Empirical Evidence. American Journal of Epidemiology. 2012; 175(8):739–49. [PubMed: 22427610]

7. Anderson CD, Nalls MA, Biffi A, Rost NS, Greenberg SM, Singleton AB, Meschia JF, Rosand J. The effect of survival bias on case-control genetic association studies of highly lethal diseases. Circulation: Cardiovascular Genetics. 2011; 4(2):188–96. [PubMed: 21292865]
8. Heikkilä K, Ebrahim S, Lawlor DA. A systematic review of the association between circulating concentrations of C reactive protein and cancer. Journal of Epidemiology & Community Health. 2007; 61:824–33. [PubMed: 17699539]
9. Meschia JF, Brott TG, Brown RD Jr, Crook RJ, Frankel M, Hardy J, Merino JG, Rich SS, Silliman S, Worrall BB. Ischemic Stroke Genetics Study. BMC Neurology. 2003; 3:4.doi: 10.1186/1471-2377-3-4 [PubMed: 12848902]
10. Pijpe A, Andrieu N, Easton DF, Kesminiene A, Cardis E, Noguès C, Gauthier-Villars M, Lasset C, Fricker JP, Peock S, Frost D, Evans DG, Eeles RA, Paterson J, Manders P, van Asperen CJ, Ausems MG, Meijers-Heijboer H, Thierry-Chef I, Hauptmann M, Goldgar D, Rookus MA, van Leeuwen FE. GENEPSO; EMBRACE; HEBON. Exposure to diagnostic radiation and risk of breast cancer among carriers of BRCA1/2 mutations: a retrospective cohort study (GENE-RAD-RISK). BMJ. 2012; 345:e5660.doi: 10.1136/bmj.e5660 [PubMed: 22956590]
11. Satten GA, Kupper L. Inferences about exposure-disease associations using probability-of-exposure information. Journal of the American Statistical Association. 1993; 88:200–8.
12. Chen J, Chatterjee N. Exploiting hardy-weinberg equilibrium for efficient screening of single SNP associations from case-control studies. Human Heredity. 2007; 63:196–204. [PubMed: 17317968]
13. Breslow N, Cain K. Logistic-regression for two-stage case-control data. Biometrika. 1988; 75:11–20.
14. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Statistics in Medicine. 1994; 13:153–62. [PubMed: 8122051]
15. Chen J, Kang G, VanderWeele T, Zhang C, Mukherjee B. Efficient designs of gene-environment interaction studies: implications of Hardy-Weinberg equilibrium and gene-environment independence. Statistics in Medicine. 2012; 31:2516–30. [PubMed: 22362617]
16. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from imcomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological). 1977; 39:1–38.
17. Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population structure in case-control studies of genetic association using a novel latent-class model. American Journal of Human Genetics. 2011; 68:466–477. [PubMed: 11170894]
18. Lin D, Weinberg CR, Feng R, Hochner H, Chen J. A Multi-Locus Likelihood Method for Assessing Parent-of-Origin Effects Using Case-Control Mother-Child Pairs. Genetic Epidemiology. 2012; 37(2):152–62. DOI: 10.1002/gepi.21700 [PubMed: 23184538]

6. Appendix

Appendix Table 1

Notational glossary.

N_1	Number of cases, some living some deceased
$\hat{N_{1j}}$	Estimated number of cases with genotype j
N_0	Number of controls, all genotyped
$\hat{N_{0j}}$	Estimated number of controls with genotype j
n_1	Number of surviving cases
m	number of deceased cases ($=N_1-n_1$)
n_{ij}	Number with case status i and genotype j
m_{jk}	Number of deceased cases with offspring genotype j and spouse genotype k
m_{jc}	Number of deceased cases with offspring genotype j and no spouse genotype

m_{jp}	Number of deceased cases with spouse genotype j
m_O	Number of deceased cases with no offspring genotype information
m_{1A}, m_{1a}	Number of deceased case for whom we have a genotyped child and who must have passed allele A (a) to the child $m_{1A} = m_{00} + m_{01} + m_{12} + m_{0c} = m_{0+} + m_{12}$; $m_{1a} = m_{10} + m_{21} + m_{22} + m_{2c} = m_{10} + m_{2+}$
m_{1Aa}	Number of deceased cases for whom we have a genotyped child and the child has genotype Aa; $m_{1Aa} = m_{1c}$
m_{1*}	Number of deceased cases whose family members' genotypes provide no information in predicting cases' genotype; $m_{1*} = m_{11} + m_O$
N_{ijk}	Number of subjects with case status i, genotype j, and environmental exposure k
Y	Disease status flag (1 if case)
G	Genotype of study participant
G^f, G^c, G^p	Genotype of deceased case's family member(s)
G^c	Genotype of deceased case's offspring
G^p	Genotype of deceased case's spouse (the other parent of that offspring)
S	Survival status flag (1 if alive)
R, R^c, R^p	Flag for recruitability of family members, offspring and spouse for deceased case (1 if recruitable)
$f(G)$	Function of G for logistic association with disease
$h(G)$	Function of G for logistic association with death, given disease
β	Genetic association parameter for disease association model
ρ	Proportion of deceased cases whose death is unrelated to G
a	Genetic association parameter for death model that conditions on disease
p_a	MAF in the control population
$\hat{p}_a^*$	Consistent estimator for the MAF in the case population
$\hat{p}_a^{**}$	Consistent estimator for the proportion of subjects with a allele in the case population
γ	Proportion of family members who are not recruitable.

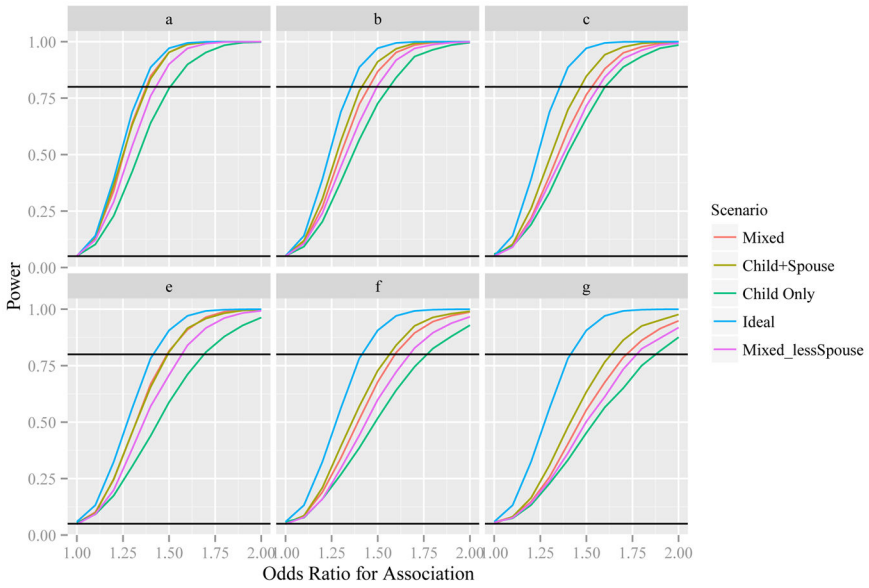

Figure 1.

Power plot.

a. Log-additive model for association and co-dominant model for survival, recruiting probability = 1. In this case, Mixed and the Child+Spouse designs are identical, and the minor difference in the figure was due to computation precision;

b. Log-additive model for association and co-dominant model for survival, recruiting probability = 0.7;

c. Log-additive model for association and co-dominant model for survival, recruiting probability = 0.5;

d. Dominant model for both association and survival, recruiting probability = 1; In this case, Mixed and the Child+Spouse designs are identical, and the minor difference in the figure was due to computation precision;

e. Dominant model for both association and survival, recruiting probability = 0.7;

f. Dominant model for both association and survival, recruiting probability = 0.5.

Stat Med. Author manuscript; available in PMC 2017 July 20.

Table 1

Estimated OR and Type I Error Rates for Testing Association (nominal level = 0.05) by FSD or when Deceased Cases were Naively Ignored.

Association Model	Survival Model	Method	True OR for Death (e^{α_1})				
			1.0	1.2	1.4	1.8	2.0
Log-additive	Co-dominant	Naïve	1.00 | 0.048	0.97 | 0.049	0.95 | 0.059	0.90 | 0.101	0.88 | 0.145
			1.00 | 0.049	0.97 | 0.051	0.95 | 0.058	0.90 | 0.084	0.88 | 0.110
		FSD	1.00 | 0.047	1.00 | 0.047	1.00 | 0.048	1.00 | 0.045	1.00 | 0.044
			1.00 | 0.048	1.00 | 0.048	1.00 | 0.047	1.00 | 0.049	1.00 | 0.048
Dominant	Dominant	Naïve	1.00 | 0.051	0.98 | 0.050	0.96 | 0.054	0.91 | 0.087	0.89 | 0.110
			1.00 | 0.049	0.98 | 0.053	0.95 | 0.057	0.91 | 0.073	0.89 | 0.087
		FSD	1.00 | 0.046	1.00 | 0.046	1.00 | 0.047	1.00 | 0.046	1.00 | 0.049
			1.00 | 0.045	1.00 | 0.046	1.00 | 0.047	1.00 | 0.047	1.00 | 0.048

* The first row corresponds to mixture proportion ρ=0.3, and the second corresponds to ρ=0.6. Before "|" is the estimate of $e^{\beta 1}$ and after is the type-I error rate for a nominally 0.05-level test of the incidence association parameter.

Stat Med. Author manuscript; available in PMC 2017 July 20.

Table 2

Parameter Estimation under Different Models and Scenarios for Genotyping Family Members of Deceased Cases.

Model	Odds Ratio[a]	γ^b	Estimates (S.D.)		
			Mixed	Child + Spouse	Child Only
Log-additive for association and co-dominant for survival	log(1) = 0	1	−0.002 (0.120)	−0.004 (0.120)	−0.005 (0.162)
		0.7	−0.002 (0.140)	−0.004 (0.132)	−0.002 (0.175)
		0.5	−0.004 (0.163)	−0.002 (0.146)	−0.009 (0.192)
		0(Naive)		−0.051 (0.130)	
	log(1.4) = 0.336	1	0.336 (0.113)	0.334 (0.113)	0.340 (0.147)
		0.7	0.335 (0.131)	0.339 (0.124)	0.335 (0.158)
		0.5	0.333 (0.151)	0.336 (0.136)	0.337 (0.172)
		0(Naive)		0.286 (0.123)	
	log(1.8) = 0.588	1	0.588 (0.109)	0.589 (0.109)	0.589 (0.139)
		0.7	0.583 (0.126)	0.588 (0.119)	0.587 (0.149)
		0.5	0.585 (0.144)	0.585 (0.131)	0.583 (0.162)
		0(Naive)		0.539 (0.120)	
Dominant for both association and survival	log(1) = 0	1	0.000 (0.143)	0.001 (0.143)	0.001 (0.193)
		0.7	0.004 (0.169)	−0.004 (0.159)	0.002 (0.210)
		0.5	−0.002 (0.197)	−0.003 (0.176)	−0.001 (0.231)
		0(Naive)		−0.046 (0.148)	
	log(1.4) = 0.336	1	0.340 (0.143)	0.337 (0.143)	0.336 (0.188)
		0.7	0.336 (0.168)	0.336 (0.158)	0.337 (0.204)
		0.5	0.338 (0.195)	0.337 (0.176)	0.338 (0.224)
		0(Naive)		0.291 (0.144)	
	log(1.8) = 0.588	1	0.590 (0.145)	0.590 (0.145)	0.591 (0.187)
		0.7	0.589 (0.170)	0.592 (0.161)	0.589 (0.204)

Model	Odds Ratio[a]	γ^b	Estimates (S.D.)		
			Mixed	Child + Spouse	Child Only
		0.5	0.592 (0.197)	0.594 (0.179)	0.586 (0.224)
		0(Naive)		0.546 (0.143)	

[a] The odds ratio of association.

[b] Recruiting probability for family members.

[*] Sample size is 500 for cases and 500 for controls; $\rho = 0.3$, $\alpha 0 = -2$, and $e^{\alpha 1} = 1.4$ (with survival probabilities 0.62, 0.59 and 0.55 for each genotype.

Table 3

OR Estimates By the Naïve and FSD methods for SNPs.

SNPs	OR Estimates			Log OR Estimates (S.E.)			P-Value		
	Naïve	FSD	FSD[f]	Naïve	FSD	FSD[f]	Naïve	FSD	FSD[f]
rs3750817	0.809	0.753	0.830	−0.213 (0.142)	−0.283 (0.153)	−0.186 (0.127)	0.1333	0.0640	0.1432
exm.rs2981579	1.341	1.518	1.323	0.294 (0.138)	0.418 (0.159)	0.280 (0.126)	0.0329	0.0085	0.0263
rs2981578	0.689	0.628	0.702	−0.373 (0.139)	−0.465 (0.161)	−0.354 (0.127)	0.0072	0.0038	0.0052
exm.rs2981575	1.382	1.457	1.379	0.324 (0.136)	0.377 (0.155)	0.321 (0.126)	0.0175	0.0154	0.0108
rs1219648	1.376	1.477	1.367	0.319 (0.137)	0.390 (0.156)	0.313 (0.126)	0.0195	0.0123	0.0133
rs2981582	1.349	1.577	1.338	0.299 (0.136)	0.456 (0.155)	0.291 (0.126)	0.0281	0.0032	0.0204

*The Naïve method utilizes the observed data only and ignores the ungenotyped cases. The FSD method assumes that cases with different genotypes (*AA, Aa* and *aa*) have different "survival" probabilities (q_0, q_1 and q_2) and incorporates genotype information of family members of ungenotyped cases. The FSD[f] method is the same as FSD except for assuming equal survival probabilities ($q_0 = q_1 = q_2 = q$).

Table 4

Recruitability ("Survival") Probabilities Estimated By FSD for SNPs.

SNPs	FSD			FSD[f]
	AA	*Aa*	*aa*	
rs3750817	0.417 (0.053)	0.425 (0.052)	0.527 (0.130)	0.433 (0.044)
exm.rs2981579	0.465 (0.088)	0.462 (0.050)	0.356 (0.067)	0.433 (0.059)
rs2981578	0.392 (0.061)	0.451 (0.051)	0.472 (0.106)	0.433 (0.050)
exm.rs2981575	0.437 (0.080)	0.454 (0.049)	0.389 (0.073)	0.433 (0.058)
rs1219648	0.432 (0.079)	0.467 (0.050)	0.366 (0.071)	0.433 (0.058)
rs2981582	0.462 (0.083)	0.474 (0.050)	0.328 (0.066)	0.433 (0.057)

[*] The Naïve method utilizes the observed data only and ignores the ungenotyped cases. The FSD method assumes that cases with different genotypes (*AA*, *Aa* and *aa*) have different "survival" probabilities (q_0, q_1 and q_2) and incorporates genotype information of family members of ungenotyped cases. The FSD[f] method is the same as FSD except for assuming equal survival probabilities ($q_0 = q_1 = q_2 = q$).