Published in final edited form as:

Biometrics. 2016 September ; 72(3): 827–834. doi:10.1111/biom.12460.

The latent class twin method

Stuart G. Baker

Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892-7354, USA (sb16i@nih.gov)

Summary

The twin method refers to the use of data from same-sex identical and fraternal twins to estimate the genetic and environmental contributions to a trait or outcome. The standard twin method is the variance component twin method that estimates heritability, the fraction of variance attributed to additive genetic inheritance. The latent class twin method estimates two quantities that are easier to interpret than heritability: the genetic prevalence, which is the fraction of persons in the genetic susceptibility latent class, and the heritability fraction, which is the fraction of persons in the genetic susceptibility latent class with the trait or outcome. We extend the latent class twin method in three important ways. First, we incorporate an additive genetic model to broaden the sensitivity analysis beyond the original autosomal dominant and recessive genetic models. Second, we specify a separate survival model to simplify computations and improve convergence. Third, we show how to easily adjust for covariates by extending the method of propensity scores from a treatment difference to zygosity. Applying the latent class twin method to data on breast cancer among Nordic twins, we estimated a genetic prevalence of 1%, a result with important implications for breast cancer prevention research.

Keywords

Breast cancer; Latent class; Missing heritability; Phantom heritability; Propensity score; Survival Variance components

1. Introduction

The twin method refers to the use of data from same-sex identical (monozygotic, MZ) and fraternal (dizygotic, DZ) twins to estimate the genetic versus environmental contribution to traits or outcomes. The standard twin method is the variance component twin method that estimates heritability, which is the fraction of the variance attributed to an additive genetic model. The variance components formulas date back one hundred years to a landmark paper by RA Fisher (1918). Investigators typically compute heritability using either a difference in correlations between MZ and DZ twins (Sham, 1998) or a mixed effects linear regression model with covariates (Kempthorne and Osborne, 1961; Jinks and Fulker 1980; Feng, Zhou, Zhang, et al, 2009). A drawback of using heritability as a measure of genetic contribution is

Supplementary Materials

A set of Mathematica packages for implementing this methodology with binary or survival outcomes is available at the Biometrics website on Wiley Online Library. The code is also available from the author upon request.

the difficulty interpreting variance components. To circumvent this drawback, Baker, Lichtenstein, Kaprio et al. (2005) formulated a latent class twin method that estimates two relevant quantities not involving variance components, namely the genetic prevalence and the heritability fraction. The genetic prevalence is the fraction of persons in the genetic susceptibility latent class. The heritability fraction is the fraction of persons in the genetic susceptibility latent class with the trait or outcome.

We extend the original latent class twin method in three important ways. First we formulate an additive genetic model in the latent class framework to broaden the sensitivity analysis beyond the original autosomal dominant and recessive genetic models. Second, we specify a separate survival model to improve convergence and computational efficiency and incorporate many age intervals without computational difficulties. Third, we show how to easily adjust for covariates when estimating the genetic prevalence and the heritability fraction by extending the method of propensity scores from a treatment differences to a function of zygosity. We discuss the latent class twin method in an analysis of Nordic twin data for breast cancer that were previously analyzed by Lichtenstein, Holm, Verkasalo, Iliadou et al. (2000) using the variance components twin method.

2. Survival Model

Each person in the Nordic twin study of breast cancer had one of three possible outcomes: observed breast cancer incidence (I^*), loss to follow-up (C^*), and death from competing risk (D^*). We start with count data rather than individual-level data to allow reproducibility of the analysis without privacy concerns. Let $(I^*, I^*), (I, C^*), (I^*, D^*), (C^*, C^*), (C^*, D^*), (D^*, D^*)$ denote age interval cross-classifications for each zygosity $z = MZ$ or DZ . We analyzed count data (Table 1), which are similar to that in Baker et al. (2005) but with more age-intervals and which assume that losses to follow-up and deaths from competing risk are noninformative censoring. These count data (Table 1) are three cross-classifications for each zygosity, $(I, I), (I, C), (C, C)$, where I denotes breast cancer incidence, C denotes noninformative censoring, and

$$(I, I) = (I^*, I^*),$$

$$(I, C) = (I^*, C^*) + (I^*, D^*),$$

$$(C, C) = (C^*, C^*) + (D^*, D^*) + (D^*, C^*). \quad (1)$$

If data are available, we suggest a sensitivity analysis that assumes death from competing risk implies cancer incidence, implying $(I, I) = (I^*, I^*) + (D^*, D^*) + (I^*, D^*)$, $(I, C) = (I^*, C^*) + (D^*, C^*)$, and $(C, C) = (C^*, C^*)$.

To transform the counts for censored survival data in Table 1 to six estimated counts for the next step in the twin analysis, we incorporate symmetries into a standard bivariate EM

algorithm for discrete survival times (Campbell, 1981; Hanley and Parnes, 1983). Let n_{zabII} , n_{zabIC} , and n_{zabCC} denote the cell counts for (I, I) , (I, C) , and (C, C) for zygosity z , where a and b index age intervals in rows and columns, respectively. We specify that censoring in an age interval implies that, in the absence of censoring, breast cancer would arise in the age interval or later, where the last age interval L includes persons who never develop breast cancer. For step j of the EM algorithm, let $e_{zab}^{(j)}$ denote the expected number of twin pairs with breast cancer incidence in age intervals a and b , except that $a = L$ or $b = L$ corresponds to no breast cancer incidence before age interval L . Because the designations of one twin in a pair as first (corresponding to a row in the table) and the other as second (corresponding to a column in the table) are arbitrary, we use the function $\text{sym}(n_{zab}) = (n_{zab} + n_{zba})/2$ to create symmetric tables in which $e_{zab}^{(j)} = e_{zba}^{(j)}$. The starting values of the EM algorithm are $e_{zab}^{(0)} = \text{sym}(n_{zabII})$ for $a < L$ and $b < L$, $e_{zab}^{(0)} = n_{zabCC}$ for $a = b = L$, and $e_{zab}^{(0)} = \text{sym}(n_{zabIC})$ otherwise. The $j+1$ step of the EM algorithm is

$$e_{zab}^{(j+1)} = \text{sym}(n_{zabII}) + \text{sym}(n_{zabIC} e_{zab}^{(j)} / \sum_{b^* \geq b} e_{zab^*}^{(j)}) + \text{sym}(n_{zabCC}) e_{zab}^{(j)} / \sum_{a^* \geq a} \sum_{b^* \geq b} e_{za^*b^*}^{(j)}. \quad (2)$$

We specify convergence as a change of less than 0.001 in $\text{Max}_{z,a,b} |e_{zab}^{(j+1)} - e_{zab}^{(j)}|$. Let e_{zab} denote the final estimated cell counts from the EM algorithm. Let $Y_{BIN} = 1$ if breast cancer incidence occurs before age interval L , and 0, otherwise. Let n_{zu} denote the estimated number of twin pairs with zygosity z and $u = 0, 1, 2$ outcomes of $Y_{BIN} = 1$. The estimated counts in Table 2 are $n_{z0} = e_{zLL}$, $n_{z1} = \sum_a \sum_b e_{zab} - n_{z0} - n_{z2}$, and $n_{z2} = \sum_{a < L} \sum_{b < L} e_{zab}$. In Section 3 we apply the variance components twin method and, in Section 4, we apply the latent class twin method to the estimated counts in Table 2.

3. Variance components twin method

To provide perspective and a basis for later formulas, we review the variance components twin method. Consider a single gene with two alleles denoted G and g . Let s and $(1-s)$ denote the frequencies of G and g . Let j index genotype pair, which we define as (genotype for twin 1, genotype for twin 2), where $j = (GG, GG), (GG, Gg), (GG, gg), (Gg, GG), (Gg, Gg), (Gg, gg), (gg, GG), (gg, Gg), (gg, gg)$. Let $f_{j|z}(s)$ denote the probability of genotype pair j given zygosity z under random mating and Hardy Weinberg equilibrium (Table 3). See Baker et al. (2005) for a derivation of $f_{j|z}(s)$. For a binary outcome, investigators typically postulate an unobserved continuous variable Y , called a liability, that corresponds to $Y_{BIN} = 1$ only when it is larger than an unobserved threshold (Pearson and Lee, 1901; Sham, 1998). Let y_{ij} denote the liability for twin $i = 1, 2$ in genotype pair j . We write $y_{ij} = y_j$ when either y_{1j} or y_{2j} would yield the same result. The mean liability for either an MZ or DZ twin is $y_{mean} = \sum_j f_{j|z}(s) y_j$.

The variances of the liabilities for MZ twins and DZ twins are

$$V_{MZ}(s) = \sum_j f_{j|MZ}(s) (y_j - y_{\text{mean}})^2, \quad (3)$$

$$V_{DZ}(s) = \sum_j f_{j|DZ}(s) (y_{1j} - y_{\text{mean}})(y_{2j} - y_{\text{mean}}). \quad (4)$$

For one MZ twin, let t_j denote the number of G alleles for genotype pair $j = (gg, gg), (Gg, Gg), (GG, GG)$, so $t_j = 0, 1, 2$, respectively. Under an additive genetic model, $Y_j = \alpha + \beta t_j + \text{error}$, the predicted liability is $y_{\text{pred}}(j) = \alpha_{\text{est}} + \beta_{\text{est}} t_j$, where α_{est} and β_{est} are the least squares estimates of α and β , respectively. Let $V_A(s) = \sum_j f_{j|MZ}(s) (y_{\text{pred}}(j) - y_{\text{mean}})^2$ and $V_D(s) = \sum_j f_{j|MZ}(s) (y_j - y_{\text{pred}}(j))^2$. The least squares estimates of α and β solve the following equations that set to zero the derivatives of the sum of squared errors with respect to α and β : $\sum_j f_{j|MZ}(s) (y_j - \alpha_{\text{est}} - \beta_{\text{est}} t_j) = 0$ and $\sum_j f_{j|MZ}(s) t_j (y_j - \alpha_{\text{est}} - \beta_{\text{est}} t_j) = 0$. Rewriting equation (3) as $V_{MZ}(s) = \sum_j f_{j|MZ}(s) (y_j - y_{\text{pred}}(j))^2 + \sum_j f_{j|MZ}(s) (y_{\text{pred}}(j) - y_{\text{mean}})^2 + 2 (y_j - y_{\text{pred}}(j)) (y_{\text{pred}}(j) - y_{\text{mean}})$ and substituting the two least squares equations into equations (3) and (4) yields, with extensive simplification via symbolic computing, the remarkable result of Fisher (1918) that $V_{MZ}(s) = V_A(s) + V_D(s)$ and $V_{DZ}(s) = \frac{1}{2} V_A(s) + \frac{1}{4} V_D(s)$. Because equations (3) and (4) also apply to sums of genes independently contributing to liability under the additive genetic model, it is convenient to write them as $V_{MZ} = V_A + V_D$, and $V_{DZ} = \frac{1}{2} V_A + \frac{1}{4} V_D$. Let V_C denote the shared environmental variance, which is the component of variance between twins in a pair that is the same for MZ and DZ twins. Let V_E denote the component of variance that is independent among twins in a pair. Let $V_{TOT:z} = V_z + V_C + V_E$ denote the total variance of liability for a twin with zygosity z . Let r_z denote the tetrachoric correlation between liabilities of twins in a pair, which is estimated from $\{n_{zu}\}$ (Pearson 1900; Sham, 1998). The variance components equations are

$$r_{MZ} = (V_{MZ} + V_C) / V_{TOT:MZ} = (V_A + V_D + V_C) / V_{TOT:MZ}, \quad (5)$$

$$r_{DZ} = (V_{DZ} + V_C) / V_{TOT:DZ} = (\frac{1}{2} V_A + \frac{1}{4} V_D + V_C) / V_{TOT:DZ}. \quad (6)$$

Heritability is $V_{MZ} / V_{TOT:MZ}$. Because it is not possible to uniquely estimate heritability from equations (5) and (6), investigators typically assume $V_D = 0$, namely the additive genetic model fits perfectly (Sham, 1998). Investigators also assume that $V_{TOT:MZ} = V_{TOT:DZ}$ (Tenesa and Haley, 2013), which is a reasonable approximation when the genetic contribution is small. Under these assumptions, heritability is $V_A / V_{TOT:MZ} = 2 (r_{MZ} - r_{DZ})$, a quantity sometimes called narrow heritability (Sham, 1998). The variance-covariance matrix in a mixed effects linear regression model for twin data arises from equations (5) and (6) with $V_D = 0$ (Feng et al, 2009).

Based on the data in Table 2, we estimated that heritability was close to zero with a 95% upper bound of 0.01. Using a mixed effects linear regression model, Lichtenstein et al. (2000) estimated heritability as 0.27. The large difference between these two heritability estimates arises from different assumptions about loss to follow-up and death from competing risk. When we assumed that losses to follow-up and deaths from competing risks imply no cancer incidence (in contrast to our assumption of noninformative censoring), we computed $n_{z0} = \sum_a \sum_b n_{zabII}$, $n_{z1} = \sum_a \sum_b n_{zabIC}$, and $n_{z2} = \sum_a \sum_b n_{zabCC}$, giving risk to counts similar to those in Lichtenstein et al. (2000) and a similar estimate of heritability, namely 0.36, as in Lichtenstein et al. (2000).

4. Latent class twin method

The latent class twin method models genetic and environmental effects as follows (with more details presented subsequently). There are two types of latent classes: an environmental susceptibility (ES) latent class and a genetic susceptibility (GS) latent class, with probabilities of occurrence depending on the model for genetic inheritance and the zygosity. The model also specifies the following characteristics of the probabilities of outcome (or trait) given the latent class. For twins in a pair in which both are in the GS latent class, each twin has an independent probability of outcome. For twins in a pair in which both are in the ES latent class, the probabilities of outcome are not restricted to independence, allowing shared environmental effects, perhaps from a condition common to both twins in a pair. For twins in a pair in which one twin is in the GS latent class and other is in the ES latent class, the twins have independent probabilities of outcome that correspond to the particular latent class.

We specify that the probabilities of outcome given the GS latent class are large (arbitrarily set to a lower bound of 0.70) for two reasons. The first reason is that often only persons with a large probability of outcome are a cost-effective target for intervention because only in that case are the benefits likely to outweigh harms. For example, individuals with a genetic marker for cancer may receive increased screening, chemoprevention, or prophylactic surgery, which is often only worthwhile if the genetic marker indicates a large probability of cancer incidence and the intervention is likely to be effective. The second reason for specifying a large probability of outcome for the GS latent class is to reflect known biology, as with hereditary cancer syndromes, which typically involve high probabilities of cancer incidence (Nagy, Sweet, and Eng, 2004).

A fortuitous consequence of modeling the GS latent class as having a large probability of outcome is that it increases the plausibility of independent probabilities of outcomes for a twin pair in the GS latent class. When the probability of outcome given GS latent class is large, any additional contribution to the probability of outcome from shared environmental effects is relatively small, making the independent outcome assumption for twins with the GS latent class a reasonable approximation. The latent class twin method also assumes that the probability cancer does not depend on zygosity given the susceptibility phenotype, a reasonable assumption if MZ and DZ twins have similar lifestyle choices related to breast cancer incidence (Baker et al, 2005).

4.1 Autosomal dominant and recessive models

We consider a single gene with two alleles denoted G and g . The GS phenotype corresponds to genotype GG under the autosomal recessive (AR) model and genotype GG or Gg under the autosomal dominant (AD) model, and the ES phenotype corresponds to the remaining genotypes. Latent class $k = 0, 1, 2$ corresponds to exactly k twins in the pair with the GS phenotype. Let subscript “*model*” index either AD or AR . Let $d_{k|j;model} = 1$ if genotype pair j contributes to latent class k , and 0 otherwise. The probability of latent class k given zygosity z is $h_{k|z;model} = \sum_j f_{j|z}(s) d_{k|j;model}$. Based on Table 3,

$$\begin{aligned} h_{2|MZ:AR} &= s^2, & h_{2|DZ:AR} &= 14s^2(1+s)^2, & h_{1|DZ:AR} &= \frac{1}{2}s^2(3-2s-s^2), \\ h_{0|MZ:AD} &= (1-s)^2, & h_{2|DZ:AD} &= 14(2-3s+s^2)^2, & h_{1|DZ:AD} &= \frac{1}{2}(1-s)^2(4-s)s. \end{aligned} \quad (7)$$

We define three parameters involving the probability of outcome given the susceptibility phenotype:

$$a = \text{pr}(Y_{BIN}=1 \text{ in one twin} \mid \text{GS phenotype}),$$

$$b = \text{pr}(Y_{BIN}=1 \text{ in both twins in a pair} \mid \text{ES phenotype}),$$

$$c = \text{pr}(Y_{BIN}=1 \text{ in exactly one twin in a pair} \mid Y_{BIN}=1 \text{ in at most one twin in the pair, ES phenotype}).$$

Let $v = \text{pr}(Y_{BIN}=1 \text{ in one twin} \mid \text{ES phenotype}) = b + (1-b)c/2$. Let $q_{u|k;model} = \text{pr}(u \text{ outcomes of } Y_{BIN}=1 \text{ in a twin pair} \mid \text{latent class } k, \text{ model})$. Under the aforementioned assumptions,

$$\begin{aligned} q_{2|2;model} &= a^2, & q_{1|2;model} &= 2(1-a)a, & q_{0|2;model} &= (1-a)^2, \\ q_{2|1;model} &= av, & q_{1|1;model} &= v(1-a) + (1-v)a, & q_{0|1;model} &= (1-a)(1-v), \\ q_{2|0;model} &= b, & q_{1|0;model} &= (1-b)c, & q_{0|0;model} &= (1-b)(1-c). \end{aligned} \quad (8)$$

Baker et al. (2005) specified a multiple gene version of the autosomal dominant and recessive models where each gene has the same allele frequency. Because the probabilities of latent classes are similar under these single and multiple gene models, we did not include a multiple gene model in our sensitivity analysis.

4.2 Additive genetic model

Under the additive model (*Add*) for a single gene with two alleles denoted G and g , we specify six latent classes indexed by $k = (\text{number of } G \text{ alleles for one twin, number of } G \text{ alleles for the other twin})$, namely, $k = (2, 2), (2, 1), (2, 0), (1, 1), (1, 0), (0, 0)$. Let $d_{klj:\text{Add}} = 1$ if genotype pair j under the additive genetic model contributes to latent class k , and 0 otherwise. The probability of latent class k given zygosity z is $h_{k/z:\text{Add}} = \sum_j f_{j/z}(s) d_{klj:\text{Add}}$. Let $a_2 = \text{pr}(Y_{BIN}=1 \text{ in one twin} \mid GG)$ and $a_1 = \text{pr}(Y_{BIN}=1 \text{ in one twin} \mid Gg)$. The additive genetic model specifies $a_1 = a_2/2$. Under the assumptions,

$$\begin{aligned} q_{2|(2,2):\text{Add}} &= a_2^2, & q_{1|(2,2):\text{Add}} &= 2(1 - a_2)a_2, & q_{0|(2,2):\text{Add}} &= (1 - a_2)^2, \\ q_{2|(2,1):\text{Add}} &= a_2 a_1, & q_{1|(2,1):\text{Add}} &= (1 - a_1)a_2 + (1 - a_1)a_1, & q_{0|(2,1):\text{Add}} &= (1 - a_2)(1 - a_1), \\ q_{2|(2,0):\text{Add}} &= a_2 \nu, & q_{1|(2,0):\text{Add}} &= \nu(1 - a_2) + (1 - \nu)a_2, & q_{0|(2,0):\text{Add}} &= (1 - a_2)(1 - \nu), \\ q_{2|(1,1):\text{Add}} &= a_1^2, & q_{1|(1,1):\text{Add}} &= 2(1 - a_1)a_1, & q_{0|(1,1):\text{Add}} &= (1 - a_1)^2, \\ q_{2|(1,0):\text{Add}} &= a_1 \nu, & q_{1|(1,0):\text{Add}} &= \nu(1 - a_1) + (1 - \nu)a_1, & q_{0|(1,0):\text{Add}} &= (1 - a_1)(1 - \nu), \\ q_{2|(0,0):\text{Add}} &= b, & q_{1|(0,0):\text{Add}} &= (1 - b)c, & q_{0|(0,0):\text{Add}} &= (1 - b)(1 - c). \end{aligned} \quad (9)$$

4.3 Measures of genetic susceptibility

Because results for MZ twin pairs apply to individuals, we write the measures of genetic susceptibility in terms of probabilities for MZ twin pairs. The genetic prevalence, which is the probability of the GS latent class, is

$$\text{GeneticPrev}_{\text{model}} = \begin{cases} h_{2|MZ:\text{model}}, & \text{for model} = AR, AD, \\ h_{(2,2)|MZ:\text{Add}} + h_{(1,1)|MZ:\text{Add}}, & \text{for model} = \text{Add}. \end{cases} \quad (10)$$

The heritability fraction, which is the probability of the GS latent class given $Y_{BIN}=1$, is

$$HF_{\text{model}} = \begin{cases} \frac{h_{2|MZ:\text{model}} a}{h_{2|MZ:\text{model}} a + h_{0|MZ:\text{model}} \nu}, & \text{for model} = AR \text{ and } AD, \\ \frac{h_{(2,2)|MZ:\text{Add}} a + h_{(1,1)|MZ:\text{Add}} a_1}{h_{(2,2)|MZ:\text{Add}} a + h_{(1,1)|MZ:\text{Add}} a_1 + h_{(0,0)|MZ:\text{Add}} \nu}, & \text{for model} = \text{Add}. \end{cases} \quad (11)$$

4.4 Estimation

We computed maximum likelihood estimates by minimizing the deviance, $2 \sum_z \sum_u n_{zu} / \sum_z \sum_u \log(n_{zu} / \sum_u n_{zu}) - 2 \sum_z \sum_u n_{zu} \log(p_{u|z:\text{model}})$, where $p_{u|z:\text{model}} = \sum_k q_{u|k:\text{model}} h_{k|z:\text{model}}$. For each value of s in a grid search, we minimized the deviance over parameters a , b , and c , while constraining a between 0.7 and 1 and b and c between 0 and 1. We first varied s coarsely. After identifying the range of s corresponding to the smallest deviances, we varied s more finely within that range. We computed bootstrap confidence intervals based on 2000 multinomial replications from the data in Table 1.

A concern in estimation was the relatively flat likelihood surface involving parameters a and s . To investigate the estimation procedure, we estimated parameters from perfect fit data created from three sets of true parameters (Table 4). We found excellent fits for parameters b and c and sufficiently good fits for parameters a and s and for genetic prevalence and heritability fraction (Table 4).

5. Incorporating covariates

Covariate data were not available for our re-analysis of the Nordic study of breast cancer. If data are available, we suggest the following method to adjust estimates for covariates that are likely related to both outcome and zygosity, such as the difference in the amount of smoking between twins in a pair. Let $w = \text{pr}(z/x)$ denote the propensity score in this context, defined as the probability of zygosity z given a vector of covariates x . A property of propensity scores (Rosenbaum and Rubin, 1984) is $\text{pr}(x/w, z) = \text{pr}(x/w)$ and $\text{pr}(z/x) = \text{pr}(z/w)$. See also Appendix A of Baker and Lindeman (2013). Consequently, $\text{pr}(w/x, z) = \text{pr}(x/z, w) \text{pr}(z/w) \text{pr}(w) / \{\text{pr}(z/x) \text{pr}(x)\} = \text{pr}(x/w) \text{pr}(z/w) \text{pr}(w) / \{\text{pr}(z/w) \text{pr}(x)\} = \text{pr}(x/w) \text{pr}(w) / \text{pr}(x) = \text{pr}(w/x)$. Let $Q(u, z)$ denote any function of the counts n_{zu} such as the genetic prevalence or the heritability fraction. Let v index quintiles of the propensity score. We can write

$$\begin{aligned}
 & \sum_x Q(u, z|x) \text{pr}(x) \\
 &= \sum_\nu \sum_x Q(u, z|x) \text{pr}(\nu|x, u, z) \text{pr}(x), \text{ by mathematical identity,} \\
 &= \sum_\nu \sum_x Q(u, z|x) \text{pr}(\nu|x, z) \text{pr}(x), \text{ because } \nu \text{ is not a function of } u, \\
 &= \sum_\nu \sum_x Q(u, z|x, \nu) \text{pr}(\nu|x, z) \text{pr}(x), \text{ because } x \text{ contains information on } \nu, \\
 &= \sum_\nu \sum_x Q(u, z|x, \nu) \text{pr}(\nu|x) \text{pr}(x), \text{ because } \text{pr}(w|x, z) = \text{pr}(w|x), \\
 &= \sum_\nu \sum_x Q(u, z|x, \nu) \text{pr}(x|\nu) \text{pr}(\nu), \text{ by mathematical identity,} \\
 &= \sum_\nu Q(u|z) \text{pr}(\nu), \text{ by mathematical identity.}
 \end{aligned} \tag{12}$$

Thus, an adjustment of $Q(u, z)$ for quintiles of the propensity score is equivalent to an adjustment of $Q(u, z)$ for many covariates, but with the advantage of less variability and generally no sparse categories.

6. Results

Fitting the latent class twin model to the estimated counts in Table 2 yielded deviances of 8.1, 8.1, and 8.5, for genetic models *AR*, *AD* and *Add*, respectively. These deviances, which

correspond to approximately the 99th percentile when resampling the perfect fit counts computed from the parameter estimates, indicated some model misspecification, but not too extreme. The estimates of the parameter a equaled the lower bound of 0.70 for models *AD* and *AR*, and equaled 0.86 for model *Add*. As a sensitivity analysis, we set the lower bound on the parameter a to zero, and obtained similar deviances and estimates of parameters, genetic prevalence and heritability fraction, except that the estimate of the parameter a was 0.46 for models *AD* and *AR*.

With regards to the three genetic models, the estimated genetic prevalence ranged from 0.00 to 0.01 with the 95th bootstrap percentile ranging from 0.02 to 0.10, and the estimated heritability fraction ranged from 0.00 to 0.01 with the 95th bootstrap percentile ranging from 0.05 to 0.010 (Table 5 and Figure 1). The estimates for genetic prevalence, but not heritability fraction, were similar to those in Baker et al (2005). A possible explanation is that the method of Baker et al (2005) may not have adequately converged to the maximum likelihood estimates due to the larger number of parameters and the nested EM algorithms. Insufficient convergence of the EM algorithm is more likely to affect the heritability fraction than the genetic prevalence because the heritability fraction depends on both a and s (which contribute to a flat likelihood surface) while the genetic prevalence depends only on s .

7. Discussion

Many investigators have criticized heritability as a measure of genetic contribution because of its sensitivity to environmental contributions and its frequent misinterpretation (Vitzthum, 2003). RA Fisher, thirty years after developing the variance components method, wrote that heritability is “one of those unfortunate short-cuts which have emerged in biometry for lack of a more thorough analysis” (Fisher, 1951). The measures of genetic contribution in the latent class twin method, namely genetic prevalence and heritability fraction, are easier to interpret than heritability. We favor the genetic prevalence over the heritability fraction because it is invariant to the environmental contributions in a population.

The estimated genetic prevalence of breast cancer of at most 1% is consistent with the sum of estimated prevalences of the two most common breast cancer susceptibility genes, *BRCA1* and *BRCA2*. The estimated prevalence of *BRCA1* ranges from 0.07% (Anglican Breast Cancer Study Group, 2000) to 0.20% (Whittemore, Gong, John, et al., 2004), and the estimated prevalence of *BRCA2* is 0.10% (Anglican Breast Cancer Study Group, 2000). The moderate to high estimated penetrance of these genes, 48% for *BRCA1* and 72% for *BRCA2* (Anglican Breast Cancer Study Group, 2000) are consistent (for *BRCA2*) with the modeling assumption of a high probability of outcome in the GS latent class.

Recent studies of breast cancer susceptibility genes (Varghese and Easton 2010; Ghoussaini, Fletcher, Michailidou, et al. 2012; Michailidou, Hall, Gonzalez-Neira, et al. 2013), which were partly motivated by the results of the Lichtenstein et al. (2000) twin study, have discovered new genetic variants with small odds ratios. (Some of these studies report a moderate contribution to familial relative risk, but this is not the same as a moderate contribution to overall breast cancer susceptibility). Based on our results we doubt that there are additional genes associated with large probabilities of breast cancer, which would make

good targets for prevention. However, the fit of the model was not as good as we would have liked, and in future work, we hope to obtain data on smoking and other covariates to determine if adjusting for these variables would improve the fit and affect the estimates. Nevertheless, given the very small estimated genetic prevalence and the increasing questioning of the role of mutations in carcinogenesis (Baker and Kaprio, 2006; Baker, 2014), we suggest that breast cancer prevention researchers focus increased effort on identifying and changing environmental risk factors for breast cancer.

The latent class twin method is applicable to a wide range of twin studies with either survival or binary outcomes where the genetic contribution of interest is characterized by a GS latent class with large independent probabilities of outcome or traits for twins in a pair.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENT

This work was supported by the National Institutes of Health.

REFERENCES

- Anglian Breast Cancer Study Group. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. *British Journal of Cancer*. 2000; 83:1301–1308. [PubMed: 11044354]
- Baker SG. A cancer theory kerfuffle can lead to new lines of research. *Journal of the National Cancer Institute*. 2015; 107:dju405. [PubMed: 25528755]
- Baker SG, Kaprio J. Common susceptibility genes for cancer: search for the end of the rainbow. *British Medical Journal*. 2006; 33:1150–1152. [PubMed: 16690677]
- Baker SG, Lichtenstein P, Kaprio J, Holm N. Genetic susceptibility to prostate, breast, and colorectal cancer among Nordic twins. *Biometrics*. 2005; 61:55–63. [PubMed: 15737078]
- Baker SG, Lindeman KL. Revisiting a discrepant result: a propensity score analysis, the paired availability design for historical controls, and a meta-analysis of randomized trials. *Journal of Causal Inference*. 2013; 1:51–82.
- Campbell G. Nonparametric bivariate estimation with randomly censored data. *Biometrika*. 1981; 68:417–422.
- Feng R, Zhou G, Zhang M, Zhang H. Analysis of twin data using SAS. *Biometrics*. 2009; 65:584–589. [PubMed: 18647295]
- Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburgh*. 1918; 52:399–433.
- Fisher RA. Limits to intensive production in animals. *British Agricultural Bulletin*. 1951; 4:217–218.
- Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, Dicks E, et al. Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nature Genetics*. 2012; 44:312–318. [PubMed: 22267197]
- Hanley JA, Parnes MN. Nonparametric estimation of a multivariate distribution in the presence of censoring. *Biometrics*. 1983; 39:129–139. [PubMed: 6871341]
- Jinks JL, Fulker DW. Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychological Bulletin*. 1970; 73:311–349. [PubMed: 5528333]
- Kemphorne O, Osborne RH. The interpretation of twin data. *American Journal of Human Genetics*. 1961; 13:320–339. [PubMed: 13752449]

- Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England Journal of Medicine*. 2000; 343:78–85. [PubMed: 10891514]
- Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci, associated with breast cancer risk. *Nature Genetics*. 2013; 45:353–361. [PubMed: 23535729]
- Nagy R, Sweet K, Eng C. Highly penetrant hereditary cancer syndromes. *Oncogene*. 2004; 23:6445–6470. [PubMed: 15322516]
- Pearson K, Lee A. Mathematical contributions to the theory of evolution. VII. On the Inheritance of characters not capable of exact quantitative measurement. *Philosophical Transactions of the Royal Society of London A*. 1900; 195:79–150.
- Sham, P. *Statistics in Human Genetics*. London: Arnold; 1998.
- Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. *Nature Reviews Genetics*. 2013; 14:139–149.
- Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*. 1984; 79:516–524.
- Varghese JS, Easton DF. Genome-wide association studies in common cancers--what have we learnt? *Current Opinion in Genetics and Development*. 2010; 20:201–209. [PubMed: 20418093]
- Vitzthum VJ. A number no greater than the sum of its parts: the use and abuse of heritability. *Human Biology*. 2003; 75:539–558. [PubMed: 14655876]
- Whittemore AS, Gong G, John EM, McGuire V, Li FP, Ostrow KL, et al. Prevalence of BRCA1 mutation carriers among U.S. non-Hispanic whites. *Cancer Epidemiology, Biomarkers, & Prevention*. 2004; 13:2078–2083.

Figure 1.
Bootstrap distribution of genetic prevalence and heritability fraction.

Table 1

Survival data on breast cancer incidence in Nordic twins. The letter I denotes breast cancer incidence. The letter C denotes noninformative censoring. The symbol - is a structural zero.

		Age interval						
Zygosity	Cross-classification	Age interval	<40	40-49	50-59	60-69	70-79	>80
MZ	(I, I)	<40	0	-	-	-	-	-
		40-49	0	2	-	-	-	-
		50-59	1	7	4	-	-	-
		60-69	0	3	2	3	-	-
		70-79	0	0	3	6	6	-
		>80	0	1	0	1	3	0
		<40	2	8	3	1	1	1
		40-49	0	22	27	18	20	9
		50-59	0	2	25	33	32	14
		60-69	1	1	5	31	56	40
	(I, C)	70-79	0	1	3	7	40	54
		>80	0	0	0	0	13	38
		<40	395	-	-	-	-	-
		40-49	28	1707	-	-	-	-
		50-59	12	29	1266	-	-	-
		60-69	12	25	94	1068	-	-
		70-79	16	25	76	273	1105	-
		>80	4	14	46	155	545	993
		<40	1	-	-	-	-	-
		40-49	1	4	-	-	-	-
DZ	(I, I)	50-59	1	1	4	-	-	-
		60-69	0	5	10	5	-	-
		70-79	0	1	5	4	3	-
		>80	0	0	0	4	2	1
		<40	5	20	10	3	3	2
		<40	5	20	10	3	3	2
		<40	5	20	10	3	3	2
		<40	5	20	10	3	3	2
		<40	5	20	10	3	3	2
		<40	5	20	10	3	3	2

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Zygosity		Cross-classification		Age interval						
				Age interval	<40	40–49	50–59	60–69	70–79	>80
				40–49	2	37	45	40	27	11
				50–59	0	2	60	72	69	32
				60–69	0	2	12	78	108	66
				70–79	0	0	9	22	89	100
				>80	0	1	3	9	18	66
		(C, C)		<40	587	-	-	-	-	-
				40–49	55	2850	-	-	-	-
				50–59	29	66	2134	-	-	-
				60–69	20	58	175	2008	-	-
				70–79	34	45	197	508	2184	-
				>80	8	34	119	382	1044	1738

Table 2

Estimated counts derived from Table 1.

Zygosity	Number of twins in a pair with breast cancer incidence before age 80		
	2	1	0
MZ	3225	649	4564
DZ	6147	1184	8020

Table 3

Probability of genotype pairs given zygosity.

<i>J</i>	Genotype	Zygosity	
		MZ	DZ
1	<i>GG, GG</i>	s^2	$\frac{1}{4} s^2 (1+s)^2$
2	<i>GG, Gg</i>	0	$\frac{1}{2} (s^2 - s^4)$
3	<i>GG, gg</i>	0	$\frac{1}{4} (s-1)^2 s^2$
4	<i>Gg, GG</i>	0	$\frac{1}{2} (s^2 - s^4)$
5	<i>Gg, Gg</i>	$2s(1-s)$	$s - 2s^3 + s^4$
6	<i>Gg, gg</i>	0	$\frac{1}{2} (2-s)(1-s)^2 s$
7	<i>gg, GG</i>	0	$\frac{1}{4} (s-1)^2 s^2$
8	<i>gg, Gg</i>	0	$\frac{1}{2} (2-s)(1-s)^2 s$
9	<i>gg, gg</i>	$(1-s)^2$	$\frac{1}{4} (2-3s+s^2)^2$

Table 4

Investigation of model fitting based on sets of perfect fit data. “True” refers to the true parameter values, which were used to generate counts.

Set	Parameters				Genetic Prevalence			Heritability fraction		
	<i>a</i>	<i>b</i>	<i>c</i>	<i>s</i>	<i>AR</i>	<i>AD</i>	<i>Add</i>	<i>AR</i>	<i>AD</i>	<i>Add</i>
1	True	0.99	0.05	0.10	0.01	0.00	0.02	0.02	0.00	0.17
	AR	0.95	0.05	0.10	0.01	0.00	---	---	0.00	---
	AD	0.84	0.05	0.09	0.01	---	0.03	---	---	0.20
	Add	0.94	0.05	0.10	0.01	---	---	0.02	---	0.10
2	True	0.80	0.10	0.20	0.05	0.00	0.10	0.10	0.01	0.31
	AR	0.87	0.10	0.20	0.05	0.00	---	---	0.01	---
	AD	0.93	0.10	0.22	0.04	---	0.07	---	---	0.26
	Add	0.98	0.10	0.21	0.04	---	---	0.07	---	0.16
3	True	0.70	0.10	0.20	0.20	0.04	0.36	0.36	0.13	0.67
	AD	0.84	0.10	0.21	0.16	0.03	---	---	0.10	---
	AR	0.75	0.10	0.24	0.17	---	0.31	---	---	0.62
	Add	0.77	0.09	0.21	0.18	---	---	0.32	---	0.52

Table 5

Estimates (5th and 95th bootstrap percentiles) based on data in Table 1. For the additive model, we estimate parameter a_2 instead of a .

Quantity	Genetic Model		
	Autosomal recessive	Autosomal dominant	Additive
Parameter a	0.70 (0.70, 1.00)	0.70 (0.70, 1.00)	0.86 (0.70, 1.00)
Parameter b	0.39 (0.37, 0.41)	0.39 (0.37, 0.41)	0.39 (0.38, 0.42)
Parameter c	0.13 (0.08, 0.14)	0.13 (0.10, 0.14)	0.12 (0.03, 0.13)
Allele frequency s	0.00 (0.00, 0.20)	0.00 (0.00, 0.01)	0.00 (0.00, 0.05)
Genetic prevalence	0.00 (0.00, 0.04)	0.00 (0.00, 0.02)	0.01 (0.01, 0.10)
Heritability fraction	0.00 (0.00, 0.07)	0.00 (0.00, 0.05)	0.01 (0.01, 0.10)