Published in final edited form as:
Genet Epidemiol. 2010 April ; 34(3): 238–245. doi:10.1002/gepi.20454.

Fitting ACE Structural Equation Models to Case-Control Family Data

K. N. Javaras[1,*], **J. I. Hudson**[2,3], and **N. M. Laird**[4]

[1] Waisman Laboratory for Brain Imaging & Behavior, University of Wisconsin-Madison, Madison, WI, USA

[2] Department of Psychiatry, Harvard Medical School, Boston, MA, USA

[3] Biological Psychiatry Laboratory, McLean Hospital, Belmont, MA, USA

[4] Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

Abstract

Investigators interested in whether a disease aggregates in families often collect case-control family data, which consist of disease status and covariate information for members of families selected via case or control probands. Here, we focus on the use of case-control family data to investigate the relative contributions to the disease of additive genetic effects (A), shared family environment (C), and unique environment (E). We describe an ACE model for binary family data; this structural equation model, which has been described previously, combines a general-family extension of the classic ACE twin model with a (possibly covariate-specific) liability-threshold model for binary outcomes. We then introduce our contribution, a likelihood-based approach to fitting the model to singly-ascertained case-control family data. The approach, which involves conditioning on the proband's disease status and also setting prevalence equal to a pre-specified value that can be estimated from the data, makes it possible to obtain valid estimates of the A, C, and E variance components from case-control (rather than only from population-based) family data. In fact, simulation experiments suggest that our approach to fitting yields approximately unbiased estimates of the A, C, and E variance components, provided that certain commonly-made assumptions hold. Further, when our approach is used to fit the ACE model to Austrian case-control family data on depression, the resulting estimate of heritability is very similar to those from previous analyses of twin data.

Keywords

variance components; heritability; path analysis; clustered binary data

1 INTRODUCTION

To study familial aggregation of a disease, investigators often sample families through case and control probands selected based on their disease status. This design is particularly efficient for rarer diseases. We refer to the resulting data, which consist of disease status and covariate information for the probands and their relatives, as *case-control family data*. The data are typically used to investigate the presence and magnitude of familial aggregation, a necessary first step for establishing the presence of genetic or shared family environmental effects [Thomas, 2004]. However, the pattern of familial aggregation in the data can

[*] Correspondence to: K. N. Javaras, 1500 Highland Avenue, Madison, WI 53705, USA. javaras@wisc.edu. Tel: (857) 928-5898.

sometimes also be used to investigate the relative contributions of additive genetic (A), shared family environmental (C), and unique environmental (E) effects. This latter goal is our focus here: more specifically, we introduce an approach that makes it possible to obtain valid estimates of the relative contributions of A, C, and E from case-control family data, assuming that certain commonly-made assumptions, including single ascertainment, hold.

Numerous existing methods can be applied to the case-control family data to first establish the presence and magnitude of familial aggregation. These methods fall into two categories: regression methods and multivariate methods. Regression methods involve fitting regression models that treat either the probands' disease statuses or the relatives' disease statuses as the outcomes. A popular example of the latter approach, which is to be preferred [Laird and Cuenco, 2003], is using logistic regression to model the probability of each relative's disease status as a function of proband disease status [Khoury et al., 1993; Hudson et al., 2001; Laird and Cuenco, 2003], with generalized estimating equations used to handle the dependence between relatives belonging to the same family [Liang and Pulver, 1996]. Multivariate methods, which in certain special cases can be effectively equivalent to regression methods, involve fitting multivariate models for the probands' and relatives' disease statuses, with some accommodation made for ascertainment based on proband disease status [Matthews et al., 2005; Zhao et al., 1998; Whittemore, 1995].

After establishing that familial aggregation is present, investigators may be interested in using the pattern of familial aggregation in the data to disentangle the relative contributions of latent genetic and environmental effects to the disease. Some of the aforementioned models for familial aggregation can be adapted for this purpose by modifying them to incorporate genealogical information in the form of relationship-pair-specific family aggregation parameters [Zhao et al., 1998]. Here, however, we focus on models specifically developed to investigate the contributions of genetic and environmental effects. These models can be formulated, equivalently, as path analysis models or as variance components models, and, further, the variance components models can be formulated and fitted within either a structural equation modeling framework or a mixed effects modeling framework [McArdle and Prescott, 2005]. The best known of these models is the classic *ACE model* for twin data, which is used to investigate genetic and environmental effects on complex (i.e., non-Mendelian) outcomes. In its variance components formulation, the ACE model is used to partition the variance in a continuous outcome of interest into variance due to **A**dditive genetic effects, to **C**ommon (or shared) family environment, and to unique (or individual) **E**nvironment. For a binary outcome of interest like disease status, the classic ACE model is combined with a liability-threshold model [Falconer 1965] such that the variance in an unobserved continuous 'liability' variable hypothesized to underlie the binary variable is partitioned into variance due to A, C, and E.

The classic ACE model for twin data can be extended to make it appropriate for more general family data. Numerous papers have proposed ACE or ACE-type models for (single outcome) binary family data, as well as approaches to fitting these models. Several papers written in the 1970s and 1980s [Morton and MacLean, 1974; Lalouel et al., 1983; Curnow and Smith, 1975] described ACE-type models for binary family data within the structural equation modeling framework; these models were developed to investigate the effects of a major gene locus, in addition to the effects of A, C, and E. More recent papers have described ACE or ACE-type models for binary family data within the mixed effects modeling framework, fitted using a likelihood-based approach [Rabe-Hesketh et al., 2008; Noh et al., 2006; McArdle and Prescott, 2005; Pawitan et al., 2004; Guo and Wang, 2002; Pfeiffer et al., 2001] or a Bayesian approach [Burton et al., 1999].

There are numerous papers that address the issue of ascertainment in pedigree analysis in general [Thompson, 1986, Chapter 8], as well as several papers, including some of the ones cited above, that address the use of non-randomly-ascertained binary family data with ACE-type models in particular. For example, Morton and MacLean [1974] (and its extension [Lalouel and Morton, 1981]) and Pfeiffer et al. [2001] describe how to adjust the likelihood for their models for certain types of non-random ascertainment. In addition, Bowden et al. [2007], Burton [2003], Glidden and Liang [2002] (including comments by Epstein [2002] and Burton [2002] and rejoinder by Glidden [2002]), Epstein [2002], and Burton et al. [2001] describe and evaluate various methods of adjusting for ascertainment when fitting genetic variance components models. Most of these papers address the scenario where *all* families (or sibships) with *k* or more affected individuals are ascertained, and most conclude that valid estimates of the variance components for the general population can be obtained from the resulting data if appropriate correction for ascertainment is made and if the underlying genetic variance components model is correct. However, to our knowledge, no paper specifically addresses the use of ACE-type models with *singly-ascertained* case-control family data (i.e., case-control family data ascertained from a population sufficiently large relative to the number of probands sampled and relative to the sizes of the families comprising the population to assume that no family is selected via more than one proband).

To fill this void, we introduce a likelihood-based approach to fitting ACE models to singly-ascertained case-control family data. The models we fit, which have been proposed previously, combine a general-family extension of the classic ACE model for twin data with a (possibly covariate-specific) liability-threshold model for binary outcomes. Our contribution is to introduce an approach to fitting the models to case-control family data that yields valid estimates of the (population) values of the A, C, and E variance components. Our approach to fitting takes its cue from Neale and Maes [2004], who note that, although using proband-ascertained data to fit variance components models can increase power to detect effects for rarer diseases, doing so requires “good information on the base rate in the population studied.” Thus, our approach to fitting involves not only adjusting the likelihood contribution of each family by conditioning on the disease status of the proband, but also fixing prevalence to a (good!) pre-specified value during fitting. In particular, for situations where the prevalence of the disease is not already well-known for the population of interest, we fix prevalence to estimates obtained from the case-control data themselves [Javaras et al., 2009].

In Section 2, we describe the case-control family data, including some necessary notation. In Section 3, we describe the ACE model for binary family data. In Section 4, we introduce our approach to fitting the model and also discuss likelihood ratio tests for comparing the fit of the full ACE model to reduced variants. In Section 5, we employ our approach to fit an ACE model to the actual data from a case-control family study of major depressive disorder (MDD) in Austria. In Section 6, we discuss the advantages and limitations of our approach, focusing on the assumptions that must hold for it to yield valid estimates. Finally, in Supplementary Materials available on the web, we describe the results of simulation experiments designed to investigate the performance of our fitting approach.

2 CASE-CONTROL FAMILY DATA

The sample consists of case-control sampled *probands* and their *relatives*, who we refer to collectively as *subjects*. There are n_a ‘affected’ probands (e.g., MDD present) and n_u ‘unaffected’ probands (e.g., MDD absent). For the sake of convenience, the families with affected probands will be indexed by $i = 1, \ldots, n_a$ and the families with unaffected probands by $i = n_a + 1, \ldots, n_a + n_u$. Family i has n_i sampled subjects; thus, the total number of sampled subjects is $n=\sum_{i=1}^{n_a+n_u} n_i$. A sampled subject from family *i* is indexed by *ij*, where $j = 1, \ldots, n_i$ and $j = 1$ if the subject is the proband. Note that our notation implies that each sampled family has only one proband and that each sampled individual belongs to only one sampled family, as would be the case in the sufficiently large population described in the Introduction.

The data for the sample contain information on the subjects' disease statuses, and possibly also on covariates such as sex and age. We let Y_{ij} denote disease status (i.e., the binary outcome) for subject *ij*: for example, $Y_{ij} = 1$ if subject *ij* has ever had MDD and $Y_{ij} = 0$ otherwise. Further, the length *n* vector *Y* contains the disease statuses for all subjects. Similarly, the length *q* vector X_{ij} and the *n* by *q* matrix **X** will refer to the *q* covariates of interest for subject *ij* and for all subjects, respectively.

Finally, the data also contain information on the familial relationships between the subjects. For the sake of simplicity, we focus here on samples that include only first-degree relatives of probands. We use $T_{ij,ij'}$ to denote the relationship between subject *ij* and subject *ij′*. Thus, $T_{i1,ij'}$ denotes the relationship of proband *i*1 to relative *ij′* and, since we restrict ourselves here to first-degree relatives, can take values ‘parent-child,’ ‘child-parent,’ or ‘sibling-sibling’ (or ‘self’ for $j' = 1$). For $j \neq 1$ and $j' \neq 1$, $T_{ij,ij'}$ denotes the relationship between relative *ij* and relative *ij′* and, for first-degree relative sampling, can take the values ‘spouse-spouse,’ ‘parent-child’ (or vice versa), ‘sibling-sibling’, ‘grandparent-grandchild’ (or vice versa), and ‘aunt/uncle-niece/nephew’ (or vice versa). We will assume that knowing $T_{i1,ij}$ and $T_{i1,ij'}$ is sufficient for determining $T_{ij,ij'}$. In other words, two relatives are related only through their proband, which disallows relationships such as double cousins.

3 ACE MODEL FOR BINARY FAMILY DATA

Here, we describe an ACE model for binary family data similar (or identical) to models described previously [e.g., Rabe-Hesketh et al., 2008; Pawitan et al., 2004; Burton et al., 1999]. As noted above, the model combines a general-family extension of the classic ACE model with a (possibly covariate-specific) liability-threshold model for binary outcomes. Further, it is formulated and fitted within the structural equation modeling framework.

In the liability-threshold model for the binary outcomes, it is assumed that

$$Y_{ij}=\begin{cases} 0 & \text{if} \quad Y^*_{ij}<t_{X_{ij}} \\ 1 & \text{if} \quad Y^*_{ij} \geq t_{X_{ij}} \end{cases}, \tag{1}$$

where Y^*_{ij} is the unobserved, continuous liability underlying the observed binary outcome Y_{ij}, and where $t_{X_{ij}}$ is a covariate-specific threshold. The covariate-specific threshold can be modeled as a linear function of covariates thought to influence disease liability [Rice et al., 1981; Chakraborty, 1986; Khoury et al., 1993, Section 7.5.2]:

$$t_{X_{ij}}=\beta X_{ij} \tag{2}$$

where β is a length *q* parameter vector quantifying how changes in X_{ij} affect $t_{X_{ij}}$. Note that the thresholds are functions of disease prevalence, as we will discuss in Section 4.

The liability is then represented as the following sum, as in the classic ACE model:

$$Y^*_{ij}=aA_{ij}+cC_{ij}+eE_{ij} \text{ for } i=1,\ldots,n_a+n_u \text{ and } j=1,\ldots,n_i, \tag{3}$$

where A_{ij}, C_{ij}, and E_{ij} are latent additive genetic, shared family environmental, and unique environmental error components, respectively, for subject *ij*. We set the means and variances of A_{ij}, C_{ij} and E_{ij}, which are arbitrary, to 0 and 1, respectively. In addition, we set the variance of Y^*_{ij} to 1. Note that the model in (3) does not include a random error term because it would be confounded with E_{ij}, which implies that any measurement error in Y_{ij} will be reflected in the unique environmental component of the model. In addition, note that the model is additive, which implies the assumption that A_{ij}, C_{ij}, and E_{ij} do not interact with each other in their effect on the liability to the disease. In addition, we assume that $\mathrm{Cov}(A_{ij}, C_{ij}) = 0$, $\mathrm{Cov}(A_{ij}, E_{ij}) = 0$, and $\mathrm{Cov}(C_{ij}, E_{ij}) = 0$. These assumptions imply, for example, that genes do not shape environment, either directly or indirectly, and they allow the variance in liability to be partitioned into separate additive genetic, shared environmental, and unique environmental *variance components*, denoted as a^2, c^2, and e^2, respectively. Since $\mathrm{var}(Y^*_{ij})=1$, these variance components can be interpreted as *proportions* of the variance in the underlying liability.

Members of the same family may have similar or even identical values for A_{ij} or C_{ij}, a fact that is reflected in the within-family correlations of the A_{ij}s and C_{ij}s, which are discussed below. If it is assumed that family members' outcomes do not directly affect each other, then the indirect effects of similar genes and shared environment on family members' liabilities are the sole source of the observed association between their outcomes, such that:

$$\mathrm{Cor}(Y^*_{ij}, Y^*_{ij'})=a^2\mathrm{Cor}(A_{ij}, A_{ij'})+c^2\mathrm{Cor}(C_{ij}, C_{ij'}). \tag{4}$$

Equation (4) means that, once the within-family correlations have been specified, the observed associations between family members' outcomes can be used to estimate a^2 and c^2 (and e^2 from $1 - a^2 - c^2$).

Specifying the within-family correlations for the additive genetic component requires several assumptions commonly made for general family data as well as for twin data. More specifically, we assume that the genetics of the disease are not influenced by dominance, epistasis, or assortative mating, which implies that $\mathrm{Cor}(A_{ij}, A_{ij'})$ equals $\frac{1}{2^{d(ij,ij')}}$, where $d(ij, ij')$ equals the degree of the relationship between *ij* and *ij′* (i.e., 1 for first-degree relatives, 2 for second-degree relatives, etc.). For spouse-spouse pairs, we assume that $d(ij, ij')$ is very large (i.e., spouses are not close relatives); this assumption, combined with that of no assortative mating, implies that $\mathrm{Cor}(A_{ij}, A_{ij'}) = 0$ for spouse-spouse pairs.

Specifying the within-family correlations for the shared environmental component is more complicated for general family data than for twin data. With twin data, it is typical to assume simply that dizygotic twins (reared together) share a family environment to the same extent as monozygotic twins (reared together). However, with more general family data, it is not plausible to assume that all types of family member pairs share a family environment to the same extent, and it can be difficult to determine the extent to which any pair of family members share a family environment. This complication can be handled in two ways. First, we could simply assume that shared environment has no effect on the outcome (i.e., $c^2 = 0$), which may be a reasonable assumption for some diseases. Second, we could attempt to measure c^2 by making certain assumptions about $\mathrm{Cor}(C_{ij}, C_{ij'})$. For instance, we could assume that Cor(C_{ij}, $C_{ij'}$) is a known function of the amount of time that *ij* and *ij′* lived together [Hopper and Matthew, 1982]. Alternatively, we could set Cor(C_{ij}, $C_{ij'}$) equal to the same (pre-specified or estimated) value for all pairs with the same $T_{ij,ij'}$ value. For example, it could be assumed that all sibling-sibling, parent-child, and spouse-spouse pairs share family environments to differing extents reflected in the correlations γ_{sib}, γ_{par}, and γ_{mar}, respectively, and that other types of relative pairs (who do not typically live together) do not share a family environment and thus have Cor(C_{ij}, $C_{ij'}$) = 0 [Thomas 2004, p. 98]. In this example, it would be necessary to impose one constraint (e.g., $\gamma_{sib} = 1$) to identify the model.

It has been noted that the assumptions of the above model may not be valid for a particular disease. As a first example, the assumptions about shared family environment may not be valid, which can result in biased estimates of the ACE variance components. As noted above, specifying the shared family environmental correlations is more difficult for general family members than for twins. In fact, for various pedigree types typically included in family data, patterns of shared family environment “easily mimic genetic transmission,” making it difficult or impossible to separate the two [Gjessing and Lie, 2008]. For instance, separating the two is difficult with the Austrian data (Section 5), as reflected in a large negative correlation between the estimates of a^2 and c^2, and may be impossible with other data (e.g., data that consist of only probands and siblings). As a second example, the assumption of no gene-environment interactions may not be entirely appropriate for some diseases [Moffitt et al., 2005]. As a third example, estimates of genetic components derived from family members who are unmatched in age may be biased downwards if different genetic factors account for the variation in liability at different ages [Maes et al., 1997]. As a fourth example, the assumptions surrounding the liability threshold model may not be valid. For instance, treating disease status as binary may not be appropriate for late-onset diseases if the data contain many young subjects whose outcomes will be effectively censored. In this case, it may be more appropriate to use an ACE model developed for survival outcomes [Pitkäniemi et al., 2007], although the ACE model for binary family data could still be used if age were incorporated as a threshold-shifting covariate. However, even if disease status can reasonably be treated as a binary outcome, some authors [e.g., Kraemer, 1997; Hopper, 1993] have questioned the validity of the liability-threshold model (but see Lyons et al. [1997] for a defense of its suitability). Relatedly, Glidden and Liang [2002] show that normal-distribution-based estimates of variance components from sibships are biased if family members’ liabilities actually follow a multivariate *t* distribution with five degrees of freedom, a point that we address further in Section 6.

4 MODEL FITTING

We take a likelihood-based approach to fitting. A likelihood for the variance components formulation of the ACE model in (3) is

$$L(a^2, c^2, \gamma, t_x|\mathbf{Y}, \mathbf{X}) \propto \prod_{i=1}^{n_a+n_u} f(Y_{i2}, \ldots, Y_{in_i}|Y_{i1}, X_{i1} \ldots, X_{in_i}), \tag{5}$$

where γ is a vector containing any shared family environmental correlations being estimated, and where t_x is a vector containing the possibly covariate-specific thresholds (the $t_{X_{ij}}$s) or else the parameters describing the relationship between the covariates and the covariate-specific thresholds (e.g., β from equation (2)). The likelihood addresses the case-control ascertainment by conditioning on the proband’s outcome, following, for example, Hopper and Matthews [1982]. It does not condition on the family’s, or the family members’, ascertainment status(es) because, as demonstrated by Tosteson et al. [1991], ascertainment status can be ignored under single ascertainment. In addition, the likelihood does not include a term modeling family size or structure alongside the term modeling family members' disease statuses, as would be done in the full likelihood approach (see Thompson [1986, Section 8.2]); this omission implies an assumption that family size and structure do not convey information about the parameters of interest (a^2 and c^2).

The conditional probabilities in (5) can be obtained from the joint probabilities of the outcomes, which can be calculated once we assume a distribution function for the Y^*_{ij}s. Here, we assume that the joint distribution of $Y^*_{i1}, \ldots, Y^*_{in_i}$ is a n_i-dimensional multivariate normal (e.g., Curnow and Smith, 1975). Thus,

$$f(Y_{i1}, Y_{i2}, \ldots, Y_{in_i} | X_{i1} \ldots, X_{in_i}) = \int_{D(Y_{in_i})} \ldots \int_{D(Y_{i1})} MVN(\mu, \sum\nolimits_i) dY^*_{i1} \ldots Y^*_{in_i} \tag{6}$$

where

$$D(Y_{ij}) = \begin{cases} [-\infty, t_{X_{ij}}] & \text{if} \quad Y_{ij}=0 \\ [t_{X_{ij}}, \infty] & \text{if} \quad Y_{ji}=1 \end{cases}, \tag{7}$$

where

$$\mu = \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix}^T, \tag{8}$$

and where

$$\sum\nolimits_{ij,ij'} = \begin{cases} 1 & \text{for} \quad j=j' \\ a^2\mathrm{Cor}(A_{ij}, A_{ij'}) + c^2\mathrm{Cor}(C_{ij}, C_{ij'}) & \text{for} \quad j \neq j' \end{cases}. \tag{9}$$

The mean vector μ is set to zero in order to identify the model because μ is completely confounded with the $t_{X_{ij}}$s. Finally, note that, in (7), the same (covariate-specific) thresholds are used for probands and for relatives, which implies an assumption that case (control) probands are randomly selected from among affected (unaffected) population members.

4.1 PARAMETER ESTIMATION

We choose to fix the threshold(s) before fitting the ACE model to the case-control family data. The reason for this choice is that the alternative approach—jointly estimating the threshold alongside a^2, c^2, and γ—produces different threshold estimates depending on the model's values for the other parameters. More specifically, for reduced variants of the ACE model that specify low or zero correlations between family members (e.g., the E model, where $a^2 = c^2 = 0$), lowering the threshold estimate improves the model's fit to data that exhibit some familial aggregation. However, since the threshold is a function of disease prevalence and is therefore theoretically the same in all models, comparisons between different variants of the ACE model should be based on how well they fit the data for the same threshold value.

In applications where there is no supplemental information on prevalence, it will need to be estimated in order to determine the threshold(s). It is possible to obtain valid estimates of prevalence from case-control family data if certain commonly-made assumptions hold (see Javaras et al. [2009]). As an example, suppose that there is a single, categorical covariate thought to affect disease liability, and suppose that subject *ij* has value $X_{ij} = x$ for that covariate. (This covariate could be the result of coarsening continuous covariates into categorical variables and/or combining multiple categorical covariates by crossing their levels.) To obtain an estimate of the prevalence corresponding to the *x* stratum, which we denote π^x, we use the following equation:

$$\widehat{\pi}^x = p_A^x \widehat{\pi} + p_U^x (1 - \widehat{\pi}), \tag{10}$$

where

$$\widehat{\pi} = \frac{p_U}{1 - p_A + P_U}, \tag{11}$$

where $p_A^x (p_U^x)$ is the proportion of case (control) probands' relatives who have covariate value *x* and are affected, and where p_A (p_U) is the proportion of case (control) probands' relatives who are affected. Then, the corresponding threshold can be obtained from $\widehat{\pi}^x$ using the following equation, which relies on the assumption (stated above) that the liabilities follow a normal distribution:

$$t_{X_{ij}} = \Phi^{-1}(1 - \widehat{\pi}^x). \tag{12}$$

The estimated threshold in (12) is approximately equal to the value obtained by jointly estimating the threshold alongside the other parameters in the true variant of the ACE model. In contrast, estimating the threshold using a reduced (e.g., $a^2 = 0$) variant of the true ACE model results in a smaller estimate of the threshold (or, equivalently, a larger estimate of the prevalence). For example, when the E model is fitted, the joint threshold estimate is approximately equal to $\Phi^{-1}(1 - p^x)$, where p^x is the proportion of (all) relatives who have covariate value *x* and are affected, a quantity that is larger than $\widehat{\pi}^x$ and upwardly biased for the true prevalence when the disease aggregates in families [Javaras et al., 2009].

Once the threshold(s) have been fixed, estimates of a^2, c^2, and γ can be obtained by maximizing (5) subject to the constraints that a^2 and c^2 are each between 0 and 1 and that $a^2 + c^2$ is less than 1. Further, elements of γ are constrained to be between −1 and 1 or, if, as in the Austrian example, it is reasonable to assume that shared family environment cannot make outcomes negatively correlated, between 0 and 1.

The simulation experiments described in the web-based Supplementary Materials suggest that the above approach to fitting yields approximately unbiased estimates of the variance components when the true variant of the ACE model is fitted. This is true even for datasets and populations as small as in the Austrian example (i.e., 64 case and 58 control probands sampled from a population of 500, 000), and the bias is even smaller for larger datasets sampled from larger populations.

4.2 INFERENCE

We can form a normal-theory-based confidence interval (CI) for a^2 (or c^2) using the standard asymptotic distribution of the parameter estimates, which is normal with mean equal to the true parameter values and variance equal to the inverted information matrix. Since the normal approximation tends to be more appropriate for quantities with an unrestricted range, we choose to form a CI for a Fisher z-transformation of ($z=0.5\ln\frac{1+a}{1-a}$) and then re-transform the upper and lower bounds to obtain a CI for a^2.

In addition, we may want to constrain the full ACE model in order to test various hypotheses, including whether the disease aggregates at all within families (Model E: $a^2 = c^2 = 0$) or whether it is affected by additive genetic effects (Model CE: $a^2 = 0$) or by shared family environment (Model AE: $c^2 = 0$). We can test these hypotheses by calculating the usual likelihood ratio test (LRT) statistic for the constrained and unconstrained models. However, the LRTs for tests such as H_0: Model AE versus H_1: Model ACE will not have the standard χ^2 asymptotic distributions because the null hypotheses constrain parameters to be on the boundary of the parameter space. P-values calculated from the standard χ^2 distribution will be conservative (i.e., too big), meaning that the standard LRT or related procedures like AIC [Akaike, 1987] will choose overly parsimonious models that result in overestimates of the retained variance components [Sullivan and Eaves, 2002]. For the hypotheses mentioned above, the true distributions of the LRT statistics are mixtures of χ^2 distributions with different degrees of freedom. The exact mixture has been derived for a number of different situations [Chernoff, 1954; Self and Liang, 1987; Stram and Lee, 1994 and 1995; Verbeke and Molenberghs, 2003], including the classic ACE model for continuous twin data [Dominicus et al., 2005], but not for the ACE model for case-control family data. Thus, we recommend using a Monte Carlo test [see Ripley, 1987] in which the p-value is determined by comparing the actual LRT statistic to the distribution of LRT values calculated from a large number of datasets simulated under the null hypothesis. Alternatively, a less time-consuming possibility is to use the p-values from the standard χ^2 distribution, but with some modification, as Dominicus et al. [2005] recommend for continuous twin data. The simulation experiments described in the Supplementary Materials suggest that, for case-control family data, using half the standard p-value works well (i.e., results in actual rejection levels approximately equal to the nominal levels) for comparing AE versus ACE, but that using the standard p-value (unhalved) works well for comparing CE versus ACE or E versus either AE or CE. The experiments also suggest that, for case-control family datasets as small as in the Austrian study, LRTs (and AIC) have limited power to detect the presence of both a^2 and c^2 if they are only small or moderate. This result is unsurprising given that Kuhnert and Do [2003] found power to be similarly limited in analogous experiments with sizeable twin datasets.

5 AUSTRIAN CASE-CONTROL FAMILY STUDY

We used our method to analyze the actual data from the Austrian study [Hudson et al., 2003]. 64 affected probands with a current DSM-IV [APA, 1994] diagnosis of MDD and 58 unaffected probands without a current or past MDD diagnosis were recruited at Innsbruck University Clinics. Adult first-degree relatives of probands were eligible for the study. Probands and relatives were interviewed using the Structured Clinical Interview for DSM-IV [First et al., 1994].

Our analyses included a total of 122 probands (one per family) and 330 first-degree relatives, all interviewed in person. Disease status was measured by a variable indicating whether the subject had been diagnosed with *lifetime* MDD (i.e., had a diagnosis of MDD at any point during their life up to the present time). Overall, 13.6% of the relatives have a lifetime diagnosis of MDD, but the estimate of overall prevalence from Equation (11) was 8.8%, with estimates of 6.0% for males and 11.3% for females from Equation (10).

We first performed preliminary analyses to investigate whether MDD aggregates in families and whether the level of aggregation differs by type of relative pair (see Table I). The overall familial aggregation odds ratio (OR) is significantly greater than 1 for both the 330 proband-relative pairs (OR = 2.7) and the 360 relative-relative pairs (OR = 5.5), indicating that MDD does aggregate in families. Turning to the relative-type-specific ORs, the aunt/ uncle - niece/nephew OR is highly significant, which suggests that genetics account for at least part of the familial aggregation of MDD if we assume that aunt/uncle - niece/nephew pairs do not typically share family environments. Further, this OR is not smaller than the parent - child and sibling - sibling ORs, which suggests that shared family environment plays a very small role, if any, in the familial aggregation of MDD. Finally, note that the spouse-spouse pairs do not contain enough information to estimate γ_{mar} when fitting the ACE models.

We fitted ACE, AE, CE, and E variants of the ACE model to the data, with thresholds fixed at values estimated from (12). The thresholds differed by sex only because a logistic regression analysis of the relatives' disease status revealed that the odds of having lifetime MDD differs significantly by sex, but not by age, in our data. In addition, for model variants that included C, γ_{sib} was fixed to one, γ_{par} was either fixed to zero or estimated, and Cor(C_{ij}, $C_{ij'}$) for all other types of relatives pairs was fixed to zero. Model fitting was performed in R using a function written by the authors. The integrals in (6) were calculated using R's pmvnorm() function, which implements the algorithms proposed by Genz [1992], and maximization of ln(L) was performed using R's optim() function, which implements the technique of Byrd et al. [1995]. For each model, the log-likelihood was unimodal, and the maximum was easily found. Standard errors were calculated from a finite difference approximation to the Hessian matrix at the maximum likelihood values, with the variance components on a Fisher-transformed scale.

Table II presents the results from fitting the different variants of the ACE model. Comparing the fit of the different variants reveals that the AE model is clearly the best-fitting model, as one might expect from examination of the relative-type-specific ORs. (In contrast, the E model is clearly the worst-fitting model, further supporting the finding that MDD aggregates in families.) The AE model fits better than the CE model with a sibling - sibling and a parent - child shared family environment, despite the fact that the latter model has one more parameter than the AE model. The AE model fits slightly worse than either of the ACE models, but the difference in $-2 \log(L)$ is only 0.3, which is far from significant according to either a Monte Carlo LRT, the standard chi-square LRT, or the chi-square LRT with the p-value halved. In the best-fitting AE model, a^2 is estimated as approximately 0.52, and the 95% confidence interval for a^2 is approximately [0.24, 0.72]. The length of this confidence interval reflects the small size of the Austrian study.

In conclusion, our analysis suggests that the familial aggregation of adult MDD can be explained almost completely by additive genetic effects, which account for approximately one half the variance in the liability to MDD. These findings are consistent with the meta-analysis performed by Sullivan et al. [2000], who found that previous studies of MDD heritability suggest a range of 0.31 – 0.42, and with the recent results of Rabe-Hesketh et al. [2008], who found that the AE model (with heritability estimated as 0.43) best fit a large twin dataset on MDD.

6 DISCUSSION

The simulation experiments and Austrian MDD application reveal that our approach to fitting ACE models makes it possible to use case-control family data to obtain valid estimates of the population (i.e., non-ascertained) ACE variance components, provided that certain commonly-made assumptions hold. This advance will enhance investigators' ability to parse genetic and environmental effects on disease—a necessary step before proceeding to molecular genetic studies—in several ways. First, using general family data instead of twin data has several advantages [Pawitan et al., 2004]. For one, unless investigators have easy access to twin registry data, case-control family data are much easier to obtain. In addition, if families have more than two members, family data contain more relative pairs, and thus more power to detect variance components, than twin data with the same number of individuals. Further, using family data makes it possible to estimate effects than cannot be estimated from standard twin data (e.g., parent-child shared family environmental effects). Second, using case-control-sampled data instead of population-sampled data offers greater power, especially for rarer diseases.

Of course, our approach is not without limitations. First, we choose to fit the models within a structural equation rather than a mixed effects modeling framework. The former does confer several advantages, namely easily interpretable parameters and straightforward adjustment for case-control sampling, but the latter is easier to implement with standard statistical software, makes it easier to incorporate covariates, and involves less computation because only three- or four-dimensional integrals need to be calculated regardless of family sizes [Rabe-Hesketh et al., 2008, p. 286]. Second, we choose to fix the threshold value(s) during model fitting, which means that the standard errors for the variance component estimates will not reflect the uncertainty surrounding the thresholds. Of course, sensitivity analysis can be performed by refitting the model using different threshold values that define a reasonable range of prevalences. Third, if the liabilities do not follow a multivariate normal distribution, estimates of the ACE variance components can be more severely biased when only certain types of families are ascertained [Glidden and Liang, 2002]. Reassuringly, though, the assumption of normal liabilities is appropriate for a complex disease resulting from the sum of many small genetic and environmental effects [Lange, 1978]. Fourth, our approach will not perform well for very rare diseases because not enough cases will be available to get reliable estimates of the variance components (or, in some cases, even the prevalence). Fifth, the validity of estimates produced by our fitting approach relies on the assumptions surrounding the case-control family sampling, which include case (control) probands being representative of affected (unaffected) population members, sampled relatives being representative of all relatives, and family size having no correlation with the disease status of its members, as well as single ascertainment. Violations of these assumptions can result in biased estimates of the ACE variance components, either via biased estimates of prevalence [Javaras et al., 2009], which in turn result in biased estimates of the variance components, or more directly (e.g., if affected relatives are less likely to be sampled, estimates of a^2 or c^2 will be too small).

These limitations aside, our approach to fitting ACE models to case-control family data performs very well when used with an actual dataset: the results from the Austrian example agree with many previous studies investigating genetic and environmental effects on MDD.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Brian D. Ripley (University of Oxford) for his advice on computational issues and Dr Barbara Mangweth-Matzek (Department of Psychiatry, Innsbruck Medical University) for access to the data from the Austrian depression study. Financial support for the first author was provided by the NIH Training Program in Psychiatric Epidemiology and Biostatistics (5 T32 MN17119-22 to KJ) and by NIH grants DK59570, P50- MH069315-04, P50-MH069315-05, P50-MH084051-01, and R01-MH43454. Financial support for the second author was provided by NIMH grant R01-MH-070377-01A.

References

Akaike H. Factor analysis and AIC. Psychometrika 1987;52:317–332.

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 4. Washington, DC: American Psychiatric Press; 1994.

Bowden J, Thompson JR, Burton PR. A two-stage approach to the correction of ascertainment bias in complex genetic studies involving variance components. Annals of Human Genetics 2006;71:220–229. [PubMed: 17354286]

Burton PR. Comment on “Ascertainment adjustment in complex diseases”. Genetic Epidemiology 2002;23:214–218. [PubMed: 12384973]

Burton PR. Correcting for nonrandom ascertainment in generalized linear mixed models (GLMMs), fitted using Gibbs sampling. Genetic Epidemiology 2003;24:24–35. [PubMed: 12508253]

Burton PR, Palmer LJ, Jacobs K, Keen KJ, Olson JM, Elston RC. Ascertainment adjustment: Where does it take us? American Journal of Human Genetics 2001;67:1505–1514. [PubMed: 11078478]

Burton PR, Tiller KJ, Gurrin LC, Cookson WOCM, Musk AW, Palmer LJ. Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. Genetic Epidemiology 1999;17:118–140. [PubMed: 10414556]

Byrd RH, Lu P, Nocedal J, Zhu C. A Limited Memory Algorithm for Bound Constrained Optimization. SIAM Journal on Scientific Computing 1995;16:1190–1208.

Chakraborty R. The inheritance of pyloric stenosis explained by a multifactorial threshold model with sex dimorphism for liability. Genetic Epidemiology 1986;3:1–15. [PubMed: 3957000]

Chernoff H. On the distribution of the likelihood ratio. Annals of Mathematical Statistics 1954;25:573–578.

Curnow RN, Smith C. Multifactorial models for familial diseases in man. Journal of the Royal Statistical Society, Series A 1975;138:131–169.

Dominicus A, Skrondal A, Gjessing HK, Pedersen NL, Palmgren J. Likelihood ratio tests in behavioral genetics: Problems and solutions. Behavior Genetics 2006;36:331–340. [PubMed: 16474914]

Epstein MP. Comment on “Ascertainment adjustment in complex diseases”. Genetic Epidemiology 2002;23:209–213. [PubMed: 12384972]

Epstein MP, Lin X, Boehnke M. Ascertainment-adjusted parameter estimates revisited. American Journal of Human Genetics 2002;70:886–895. [PubMed: 11880949]

Falconer DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. Annals of Human Genetics 1965;29:51–76.

First, MB.; Spitzer, RL.; Gibbons, M.; Williams, JBW. Structured Clinical Interview for Axis I DSM-IV Disorders - Patient Edition (SCID-I/P, version 2.0). New York: Biometrics Research Department, New York State Psychiatric Institute; 1994.

Genz A. Numerical computation of multivariate normal probabilities. Journal of Computational and Graphical Statistics 1992;1:141–149.

Gjessing HK, Lie RT. Biometrical modelling in genetics: are complex traits too complex? Statistical Methods in Medical Research 2008;17:75–96. [PubMed: 17855744]

Glidden DP. Rejoinder on “Ascertainment adjustment in complex diseases”. Genetic Epidemiology 2002;23:219–220. [PubMed: 12384974]

Glidden DV, Liang KY. Ascertainment adjustment in complex diseases. Genetic Epidemiology 2002;23:201–208. [PubMed: 12384971]

Guo G, Wang J. The mixed or multilevel model for behavior genetic analysis. Behavior Genetics 2002;32:37–49. [PubMed: 11958541]

Hopper JL. Variance components for statistical genetics: applications in medical research to characteristics related to human diseases and health. Statistical Methods in Medical Research 1993;2:199–223. [PubMed: 8261258]

Hopper JL, Matthews JD. Extensions to multivariate normal models for pedigree analysis. Annals of Human Genetics 1982;46:373–383. [PubMed: 6961886]

Hudson JI, Laird NM, Betensky RA. Multivariate logistic regression for familial aggregation of two disorders: I. Development of models and methods. American Journal of Epidemiology 2001;153:501–505.

Hudson JI, Mangweth B, Pope HG Jr, De Col C, Hausmann A, Gutweniger S, Laird NM, Biebl W, Tsuang MT. Family study of affective spectrum disorder. Archives of General Psychiatry 2003;60:170–177. [PubMed: 12578434]

Javaras KN, Laird NM, Hudson JI, Ripley BD. Estimating disease prevalence using relatives of case and control probands. Biometrics. 2009 in press.

Khoury, MJ.; Beaty, TH.; Cohen, BH. Fundamentals of Genetic Epidemiology. Oxford: Oxford University Press; 1993.

Kraemer HC. What is the ‘right’ statistical measure of twin concordance (or diagnostic reliability and validity)? [Commentary]. Archives of General Psychiatry 1997;54:1121–1124. [PubMed: 9400348]

Kuhnert PM, Do KA. Fitting genetic models to twin data with binary and observed categorical responses: A comparison of structural equation model and Bayesian hierarchical models. Behavior Genetics 2003;33:441–454. [PubMed: 14574143]

Laird NM, Cuenco KT. Regression methods for assessing familial aggregation of disease. Statistics in Medicine 2003;22:1447–1455. [PubMed: 12704608]

Lalouel JM, Morton NE. Complex segregation analysis with pointers. Human Heredity 1981;31:312–321. [PubMed: 7333620]

Lalouel JM, Rao DC, Morton NE, Elston RC. A unified model for complex segregation analysis. American Journal of Human Genetics 1983;35:816–826. [PubMed: 6614001]

Lange K. Central limit theorems of pedigrees. Journal of Mathematical Biology 1978;6:59–66.

Liang KY, Pulver AE. Analysis of case-control/family sampling design. Genetic Epidemiology 1996;13:253–270. [PubMed: 8797008]

Lyons MJ, Faraone SV, Tsuang MT, Goldberg J, Ramakrishnan V, Eaves LJ, Meyer J, True WR, Eisen SA. Another view on the ‘right’ statistical measure of twin concordance [Commentary]. Archives of General Psychiatry 1997;54:1126–1128. [PubMed: 9400349]

Maes HMM, Neale MC, Eaves LJ. Genetic and environmental factors in relative body weight and human adiposity. Behavior Genetics 1997;27:325–351. [PubMed: 9519560]

Matthews AG, Finkelstein DM, Betensky RA. Analysis of familial aggregation in the presence of varying family sizes. Applied Statistics 2005;54:847–862.

McArdle JJ, Prescott CA. Mixed-effects variance components models for biometric family analysis. Behavior Genetics 2005;35:631–652. [PubMed: 16184490]

Moffitt TE, Caspi A, Rutter M. Strategy for investigating interactions between measured genes and measured environments. Archives of General Psychiatry 2005;62:473–481. [PubMed: 15867100]

Morton NE, MacLean CJ. Analysis of family resemblance. III. Complex segregation of quantitative traits. American Journal of Human Genetics 1974;26:489–503. [PubMed: 4842773]

Neale, MC.; Maes, HHM. Methodology for Genetic Studies of Twins and Families. Dordrecht: Kluwer Academic Publishers; 2004.

Noh M, Yip B, Lee Y, Pawitan Y. Multicomponent Variance Estimation for Binary Traits in Family-Based Studies. Genetic Epidemiology 2006;30:37–47. [PubMed: 16265627]

Pawitan Y, Reilly M, Nilsson E, Cnattingius S, Lichtenstein P. Estimation of genetic and environmental factors for binary traits using family data. Statistics in Medicine 2004;23:449–465. [PubMed: 14748038]

Pfeiffer RM, Gail MH, Pee D. Inference for covariates that accounts for ascertainment and random genetic effects in family studies. Biometrika 2001;88:933–948.

Pitkäniemi J, Moltchanova E, Haapala L, Harjutsalo V, Tuomilehto J, Hakulinen T. Genetic random effects model for family data with long-term survivors: Analysis of Diabetic Nephropathy in Type 1 Diabetes. Genetic Epidemiology 2007;31:697–708. [PubMed: 17487884]

Rabe-Hesketh S, Skrondal A, Gjessing HK. Biometrical modeling of twin and family data using standard mixed model software. Biometrics 2008;64:280–288. [PubMed: 17484777]

Rice J, Nichols PL, Gottesman II. Assessment of sex differences for multifactorial traits using path analysis: application to learning difficulties. Psychiatry Research 1981;4:301–312. [PubMed: 6943596]

Ripley, BD. Stochastic Simulation. New York: Wiley; 1987.

Self SG, Liang KY. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 1987;82:605–610.

Sullivan PF, Eaves LJ. Evaluation of analyses of univariate discrete twin data. Behavior Genetics 2002;32:221–227. [PubMed: 12141783]

Sullivan PF, Neale MC, Kendler KS. Genetic epidemiology of major depression: review and meta-analysis. American Journal of Psychiatry 2000;157:1552–1562. [PubMed: 11007705]

Thomas, DC. Statistical Methods in Genetic Epidemiology. Oxford: Oxford University Press; 2004.

Thompson, EA. Pedigree Analysis in Human Genetics. Baltimore, MD: The Johns Hopkins University Press; 1986.

Tosteson TD, Rosner B, Redline S. Logistic regression for clustered binary data in proband studies with application to familial aggregation of sleep disorders. Biometrics 1991;47:1257–1265. [PubMed: 1786318]

Verbeke G, Molenberghs G. The use of score tests for inference on variance components. Biometrics 2003;59:254–262. [PubMed: 12926710]

Whittemore AS. Logistic regression of family data from case-control studies. Biometrika 1995;82:57–67.

Zhao LP, Hsu L, Holte S, Chen Y, Quiaoit F, Prentice RL. Combined association and aggregation analysis of data from case-control family studies. Biometrika 1998;85:299–315.

Table I

Unadjusted Familial Aggregation Odds Ratios (OR) for the Austrian MDD Data

	Number of Pairs	Familial Aggregation OR[a]
All Pairs		
Proband-Relative Pairs	330	2.7**
Relative-Relative Pairs	360	5.1***
Sibling-Sibling Pairs		
Proband-Relative Pairs	144	2.5*
Relative-Relative Pairs	166	7.1***,[b]
Parent-Child Pairs		
Proband-Relative Pairs	186	3.4*
Relative-Relative Pairs	92	3.0
Spouse - Spouse Paris		
Proband-Relative Pairs	0	NA[c]
Relative-Relative Pairs	30	NA[c]
Grandparent - Grandchild Pairs		
Proband-Relative Pairs	0	NA[c]
Relative-Relative Pairs	3	NA[c]
Aunt/Uncle - Niece/Nephew Pairs		
Proband-Relative Pairs	0	NA[c]
Relative-Relative Pairs	69	7.5**

*Significant at 0.05;

**Significant at 0.01;

***Significant at 0.001

[a] OR = Odds(MDD | Relative has MDD)/Odds(MDD | Relative does not have MDD)

[b] Based on 332 ordered sibling-sibling pairs.

[c] Cannot estimate OR due to small sample size or small number affected.

Table II

Results from Fitting ACE Variants to the Austrian MDD Data

Model	Estimates		−2 ln(*L*)
	Cor(C_{ij}, $C_{ij'}$)[a]	Variance Components	
1: ACE	$\gamma_{sib} = 1$[b] $\hat{\gamma}_{par} = 0$	$\widehat{a^2}=0.443$ $\widehat{c^2}=0.067$ $\widehat{e^2}=0.490$	240.8
2: ACE	$\gamma_{sib} = 1$[b] $\gamma_{par} = 0$[b]	$\widehat{a^2}=0.443$ $\widehat{c^2}=0.067$ $\widehat{e^2}=0.490$	240.8
3: AE	γ_{sib} =NA γ_{par} =NA	$\widehat{a^2}=0.517$ $c^2 = 0$[b] $\widehat{e^2}=0.483$	241.1
4: CE	$\gamma_{sib} = 1$[b] $\hat{\gamma}_{par} = 0.672$	$a^2 = 0$[b] $\widehat{c^2}=0.293$ $\widehat{e^2}=0.707$	241.9
5: CE	$\gamma_{sib} = 1$[b] $\gamma_{par} = 0$[b]	$a^2 = 0$[b] $\widehat{c^2}=0.297$ $\widehat{e^2}=0.703$	246.3
6: E	γ_{sib} =NA γ_{par} =NA	$a^2 = 0$[b] $c^2 = 0$[b] $e^2 = 1$[b]	264.2

[a] Shared family environmental correlation assumed to equal zero for all spouse-spouse, grandparent-grandchild, and aunt/uncle - niece/nephew pairs.

[b] Value fixed rather than estimated.