

On the Validity of the TDT Test in the Presence of Comorbidity and Ascertainment Bias

James M. Robins,^{1*} Jordan W. Smoller,² and Kathryn L. Lunetta³

¹*Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, Massachusetts*

²*Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts*

³*Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts*

Comorbidity, the association of two disorders, occurs commonly with complex diseases. In this paper, we investigate the effects of both true (within-family) comorbidity and spurious comorbidity due to ascertainment bias on the validity of both the parental and sibling control transmission/disequilibrium test. Specifically, we consider settings in which a candidate gene is unlinked to the target phenotype but is in linkage disequilibrium with a comorbid phenotype. We derive conditions under which the presence of true and/or spurious comorbidity will result in an artificial correlation between the target phenotype and the candidate gene. *Genet. Epidemiol.* 21:326–336, 2001. © 2001 Wiley-Liss, Inc.

Key words: ascertainment bias; causal directed acyclic graphs; comorbidity; transmission/disequilibrium test

INTRODUCTION

Comorbidity, the association of two disorders, occurs commonly with complex diseases. For instance, in population surveys, the estimated lifetime prevalence of depression among individuals with social phobia ranges from 17 to 37% [Schneier et

Contract grant sponsor: National Institutes of Health; Contract grants: AI32475-09 and MH59532-02.

*Correspondence to: James M. Robins, M.D., Professor of Epidemiology and Biostatistics, Departments of Epidemiology and Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115. E-mail: robins@hsph.harvard.edu

Received for publication 5 June 2000; revision accepted 12 January 2001

© 2001 Wiley-Liss, Inc.

al., 1992; Magee et al., 1996] but is less than 15% in those without social phobia, possibly because social phobia leads to depression or the two disorders have underlying common causes. Similarly, hyperlipidemia and coronary heart disease (CAD) are associated in population surveys because hyperlipidemia is a cause of CAD and because they may have other common causes. Furthermore, it is frequently observed that two disorders are even more highly associated in clinical samples than in population surveys. The explanation for this difference is that individuals with undiagnosed social phobia or hyperlipidemia are more likely to receive medical attention and be diagnosed (i.e., ascertained) if they also have depression or CAD, respectively. Thus, even were two disorders independently distributed in the population, nonetheless, in clinical samples, they may be associated. We refer to the component of the association between two disorders that is attributable to such ascertainment bias as spurious comorbidity.

The main purpose of this paper is to study the effects of both true (within-family) comorbidity and spurious comorbidity due to ascertainment bias on our ability to test for a genetic basis of a target disease or phenotype D (e.g., social phobia or hyperlipidemia) when a comorbid disease or phenotype C (e.g., depression or CAD) itself has a genetic basis. In particular, we will study the effects of true and spurious comorbidity on the validity of both 1) the case-parental transmission/disequilibrium test (TDT) developed by Terwillinger and Ott [1992] and Spielman et al. [1993] and 2) the case-sibling control TDT (sib TDT) studied by Curtis [1997], Spielman and Ewens [1998], Boehnke and Langefeld [1998], and Horvath and Laird [1998]. We study these tests both in the setting in which data on the comorbid illness (e.g., depression or CAD) are available and the setting in which such data are missing. More specifically, consider a candidate gene G that either is a contributing cause of the comorbid disease C or is in linkage disequilibrium with such a gene but that is in linkage equilibrium with any gene affecting the target disease D . Our goal is to determine settings in which the presence of true and/or spurious comorbidity will result in an artificial correlation between the target disease D and the candidate gene G . We restrict attention to within-family designs. However, biases analogous to those we find for within-family designs will also occur in standard case-control studies based on clinically ascertained cases and unrelated controls.

A second purpose of this paper is to introduce the genetic epidemiologic community to the usefulness of causal directed acyclic graphs (DAGs) as an analytic aid. Causal DAGs were developed by Pearl and Verma [1991] and Spirtes et al. [1993] and introduced to the statistical community in Pearl [1995] and to the general epidemiologic community in Greenland et al. [1998] and Robins [2001]. Graphic models, including DAGs, have previously appeared in the genetic epidemiology literature [e.g., see Lange and Elston, 1975]. However, causal DAGs differ from purely statistical DAGs in that they encode causal (as well as probabilistic relations) in a manner that is totally rigorous and yet intuitive and easily visualized. More precisely, a causal DAG is a graphic representation of a recursive nonparametric structural equation model (RNPSEM) [Pearl, 1995]. RNPSEMs are a nonparametric nonlinear generalization of recursive linear structural equation models [Pearl, 2000] and are closely related to the finest fully randomized causally interpreted structured tree graphs of Robins [1986]. In contrast to recursive linear structural equation models, RNPSEMs allow both for discrete response variables and for arbitrary (i.e., nonparametric) nonlinear regression

of a response variable on its determinants. Both graphic models based on DAGs and recursive linear structural equation models have close connections with Sewall Wright's path models [Cox and Wermuth, 1993; Pearl, 2000]. However, in contrast with path models, because of the possible nonlinearity of the regression, the strength of an arrow on a causal DAG cannot be quantified by a path coefficient.

A CAUSAL MODEL

In this section, we describe the formal causal model on which our results are based. Our causal model is encoded by the DAG 1 in Fig. 1A representing within-family associations. The vertices on the graph represent random variables. The variable D takes value 1 if a subject truly has the target disease regardless of whether he/she has been diagnosed; $D = 0$ otherwise. Similarly, $C = 1$ if a subject has the comorbid condition and is 0 otherwise. The ascertainment variable A takes value 1 if a subject has been diagnosed with (i.e., ascertained to have) condition D and $A = 0$ otherwise.

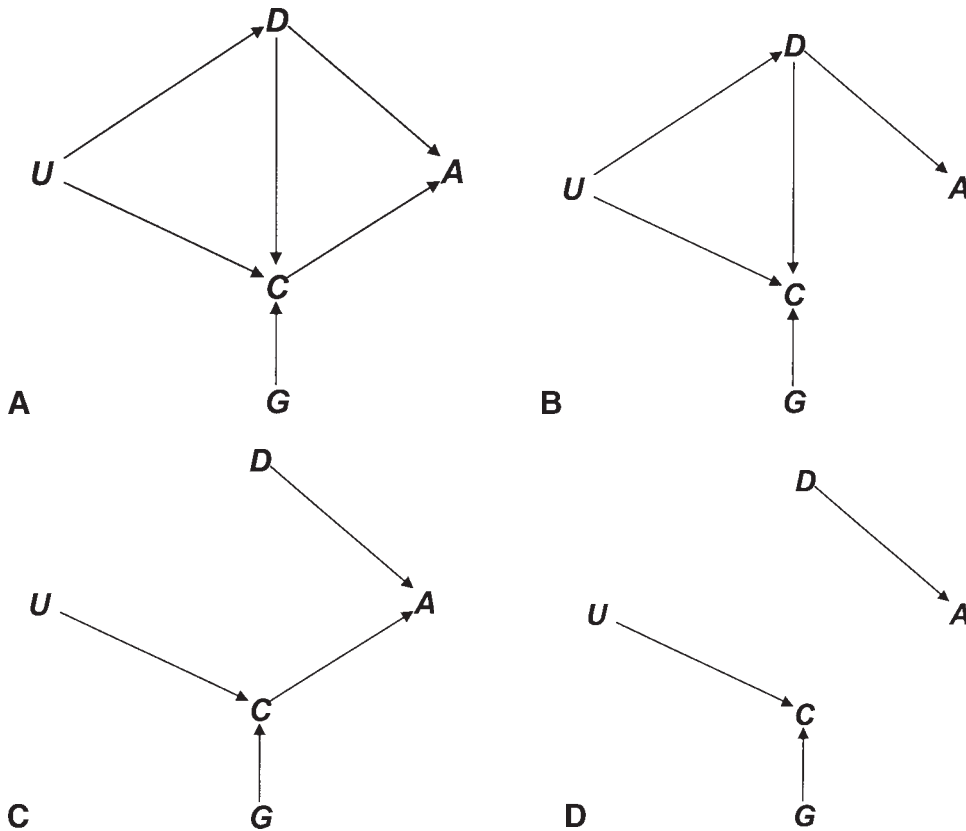


Fig. 1. Candidate causal DAGs to represent the data from a family-based case-control study with potential for both the true and spurious comorbidity. **A:** DAG 1, both true and spurious comorbidity; **B:** DAG 2, true comorbidity only; **C:** DAG 3, spurious comorbidity only; **D:** DAG 4, neither true nor spurious comorbidity. G , candidate gene; C , comorbid disease; D , target disease; A , ascertainment of target disease; U , unmeasured potential common cause of target and comorbid disease.

G represents a biallelic autosomal candidate gene; its value indicates the number of copies of a putative susceptibility allele. Finally, the variable U represents an unmeasured and unknown potential common genetical or environmental cause of both the target disease D and the comorbid disease C . For simplicity, we assume that our diagnostic procedures have 100% sensitivity and specificity. So, for example, every subject with the target condition D who receives a diagnostic test for that condition will be successfully ascertained to have it. However, there may be many persons with conditions D and/or C who have not come to medical attention and thus remain undiagnosed.

The arrows on the DAG 1 are interpreted as follows. The arrow from the candidate gene G to the comorbid disease C indicates it is either a contributing cause of or in linkage disequilibrium with a contributing cause of that disease. The lack of an arrow from G to the target disease D indicates that it is neither a contributing cause nor is in linkage disequilibrium with a contributing cause of that disease. The arrows from U to the diseases D and C indicate that it is a common cause of these diseases. The arrows from D and C to A indicate that both the target and comorbid disease are causes of having the target disease clinically diagnosed. The arrow from D to C indicates that the target disease can cause the comorbid disease C . For example, social phobia causes individuals to become depressed and hyperlipidemia causes CAD. The absence of an arrow from C to D encodes the assumption that the comorbid disease is not a cause of the target disease. In the Discussion section, we reconsider this assumption. The absence of an arrow between U and G on the graph encodes the assumption that within families, U and G are uncorrelated. For example, U could represent an unknown gene that causes both the target and comorbid disease that is unlinked to the candidate gene G . The causal relationships encoded in DAG 1 include both true and spurious comorbidity. In contrast, the causal relationships encoded in DAG 2 include true but not spurious comorbidity. Those in DAG 3 include spurious but not true comorbidity. Finally, those in DAG 4 imply that both true and spurious comorbidity are absent. Formally, we make the following definitions.

Definition: Spurious comorbidity exists if, as in DAGs 1 and 3, there is an arrow from C to A and thus ascertainment bias is present.

Definition: True (within-family) comorbidity exists if, as in DAGs 1 and 2, either 1) the arrow from D to C is present and/or 2) both the arrow from U to D and from U to C are present, and thus, within families, the two illnesses are associated.

To link the topology of our causal DAGs to the statistical data obtained in the genetic-epidemiologic study, we need to endow our DAGs with a statistical model. That is, we need to specify what restrictions the topology of a causal DAG places on the joint distribution of the variables on the graph. To do so, we let $V = (V_1, \dots, V_m)$ represent the variables on a generic causal DAG. For any variable V_i , let PA_i denote the set of variables that are parents of V_i on the graph (i.e., those variables with arrows pointing directly into V_i). We refer to V_i as a child of any of its parent nodes. Let DE_i denote the set of variables on the graph that are descendants of V_i , i.e., the variables that can be reached starting from V_i by following a sequence of directed arrows pointing away from V_i . For example, on DAG 1, A is a descendant of U but G is not. Let AN_i denote the variables on the graph that are ancestors of V_i , i.e., those variables that have V_i as a descendant. Then we make the following assumption.

Assumption A

Each variable V_i on a DAG is statistically independent of its non-descendants conditional on its parents PA_i .

Note that the descendants of a variable V_i are precisely those variables that are causally affected by V_i (or by a gene in linkage disequilibrium with V_i) either directly or indirectly through other variables. Further, note that the parents of V_i are the variables that directly cause V_i . It follows that assumption A encodes the intuitive notion that, conditional on the direct causes of a variable V_i , any variable that is not caused by V_i will be conditionally independent of V_i . Spirtes et al. [1993] refer to this assumption as the causal Markov assumption.

Example

On DAG 1, U is independent of A given D and C because U is not a descendant of A and (D, C) are the parents of A .

It turns out that this assumption logically implies additional conditional and unconditional statistical independencies. The d-separation criteria of Pearl [1995] and Lauritzen et al. [1990] show how to obtain all conditional and unconditional independencies implied by our assumption A. Specifically, our assumption implies that a set of variables X is conditionally independent of another set of variables Y given a third set of variables Z if X is d-separated from Y given Z on the graph, where d-separation, described below, is a statement about the topology of the graph. To describe d-separation, we first need to define the moralized ancestral graph generated by the variables in X , Y , and Z . In the following, a path between two variables (nodes) is any unbroken sequence of edges (regardless of the directions of the arrows) connecting the two nodes.

Definition [Lauritzen et al., 1990]

The moralized ancestral graph generated by the variables in X , Y , and Z is formed as follows.

1. First remove from the DAG all nodes (and corresponding edges) except those contained in the sets X , Y , and Z and their ancestors.
2. Next, connect by an edge every pair of nodes that share a common child.

Definition

X and Y are d-separated given Z if and only if on the moralized ancestral graph generated by X , Y , and Z , some node in Z intercepts (i.e., lies on) any path between any node in X and any node in Y .

Example 1

Consider Fig. 2 in which X is an environmental cause and Y is a genetic cause of a phenotype D . As is often true of genetic and environmental causes, X and Y are distributed independently. This follows either from assumption A or from the fact that X is d-separated from Y given Z equal to the empty set: In step 1 of the moralized ancestral graph algorithm, D (and the arrows pointing into it) is removed from the graph so that in step 2, X and Y have no children. Thus, there is no path linking X and Y on the moralized ancestral graph, so they are d-separated. In contrast, X and Y are not d-separated given D . To see this, note that on identifying Z as D , D is no

longer removed in step 1 of the moralization algorithm. Hence, in step 2, X and Y have to be connected by an edge because they have a common child D . Hence, X and Y are not d-separated given D because there is an edge between them in the moralized ancestral graph that does not intercept D . This example tells us that if we condition on a common effect D of two independent causes X and Y , we render those causes conditionally dependent. For example, if we know a subject has the phenotype D but does not have the environmental cause X , then it becomes likely that he has the genetic cause Y (because we require some explanation for the phenotype).

Here are some further examples that will be important later.

Example 2

On DAGs 1 and 3 in Figs. 1A and C, G is not d-separated from A with Z equal to the empty set because of the path $G - C - A$ on the moralized ancestral graph. But, on DAG 3, G is d-separated from A given C because on the moralized ancestral graph, the only path between G and A is intercepted by $Z = C$. However, neither on DAG 1 or DAG 2 is G d-separated from A given C . This is because on these graphs (in contrast to DAG 3), U and D are ancestors of C . Therefore, in step 2 of the moralization algorithm we add edges between G and D and between G and U . As a consequence, on the moralized ancestral graph, the path $G - D - A$ is not intercepted by $Z = C$ and thus d-separation fails.

In addition to assumption A, we make one further assumption.

Assumption B

The distribution of the variables on the graph generating the data has no additional conditional and unconditional independencies beyond those implied, through the d-separation criteria, by the topology of the graph. In the nomenclature of Spirtes et al. [1993], the statistical distribution of the data is faithful to the graph.

Here is the intuition behind assumption B. Any independencies not explained by the causal structure of the graph can be considered “accidental,” resulting from the exact balancing of positive and negative causal effects. For example, in Fig. 1, U is a parent of and thus not d-separated from C . Yet if the direct effect of U on C is equal in magnitude but opposite in direction to the effect of U on C mediated through the variable D , U and C would be independent. Because such precise fortuitous balancing is highly unlikely to occur in reality, we eliminated the possibility by imposing assumption B.

Assumptions A and B taken together imply that X and Y are d-separated given Z if and only if X and Y are statistically independent given Z .

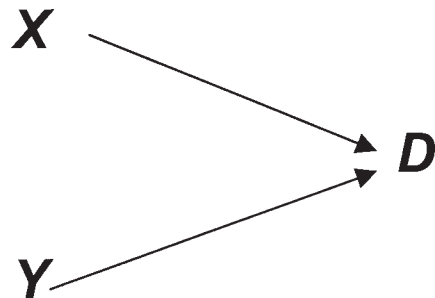


Fig. 2. A causal DAG representing independent causes of a common effect. D , phenotype of interest; X , environmental cause of D ; Y , genetic cause of D .

We are now ready to determine the conditions under which the TDT and/or sib TDTs will incorrectly conclude that gene G is associated with the target disease D . We examine the sib TDT first. Without loss of generality, we restrict attention to the simplest possible version of the test.

However, before proceeding we discuss one philosophical point. Many statisticians and some epidemiologists disparage causal language, arguing that data can only tell us about associations and not causation. We do not have the space here to enter into this seemingly never-ending argument. We only say to those who dislike causal language that all the results in this paper are mathematical consequences of assumptions A and B, and these assumptions concern purely statistical relationships between variables. We only used causal language and reasoning as motivation. So the reader is free to simply read this paper as deriving the logical consequences of the statistical assumptions A and B.

THE SIB TDT TEST

Sib TDT

We will use the terminology sib TDT to refer to any test using cases and their unaffected sib controls to test whether A (the diagnosed target disease) is associated with the candidate gene G . In the very simplest version of the sib TDT, we 1) find all clinically ascertained subjects affected with target disease D in a given catchment area with at least one undiagnosed sib and select a single undiagnosed sib at random to serve as a control and 2) score subjects as exposed if they have one or more copies of the putative susceptibility allele and otherwise call them unexposed.

To keep exposition simple, and without loss of generality, we suppose the exposed have one copy of the susceptibility allele and the unexposed have 0. In this case, the sib TDT is the McNemar test applied to the sib pairs who are discordant at the candidate gene locus. Specifically, let b be the number of discordant pairs with the affected sib exposed and c be the number of discordant pairs in which the affected sib is unexposed. The McNemar test tests the null hypothesis that conditional on $b + c$, b has a binomial $(b + c, 1/2)$ distribution.

If the nominal α -level sib TDT has an actual rejection rate of greater than α under the null hypothesis (represented in DAGs 1–4 in Fig. 1) that G neither causes nor is linked to a cause of D , we will say that the sib TDT is invalid. When the test is invalid, it will tend to find an artifactual significant association between the candidate gene G and the target disease D . In fact, as the sample size approaches infinity, the probability of finding an artifactually significant association goes to 1. Our first result is stated in the following propositions.

Proposition 1: Suppose assumptions A and B hold. If spurious comorbidity exists (i.e., DAG 1 or 3 generated the data), then the sib TDT test is invalid. If there is no spurious comorbidity (i.e., either DAG 2 or DAG 4 generated the data), then the sib TDT test is valid even in the presence of true within-family comorbidity.

Proof: Under our sib-pair case-control design, the sib TDT simply tests whether A (diagnosed target disease) is independent of genotype G . If there is spurious comorbidity (i.e., an arrow from C to A), then, as argued in example 2, G

is not d-separated from A . Conversely, if there is no spurious comorbidity (i.e., no arrow from C to A), G is independent of A because A is a non-descendant of G .

The above result assumes that data on the comorbid condition C were unavailable or not used for data analysis. If data on C status are available, we can consider selecting unaffected control siblings for each affected proband who are matched to the proband on the comorbid phenotype C . Specifically, the matched sib TDT test differs from the unmatched only in that control siblings are matched to the proband on the comorbid phenotype. If a proband has no unaffected sibling with the same comorbid phenotype, that proband is excluded from the analysis. We then have the following proposition.

Proposition 2: Suppose assumptions A and B hold. If there exists true within-family comorbidity (whether due to an unmeasured common cause U or to D causing C or to both), then the matched sib TDT is invalid, even in the absence of spurious comorbidity. In particular, the test is invalid if either DAG 1 or DAG 2 generated the data. If true within-family comorbidity is absent, then the matched sib TDT test is valid even in the presence of spurious comorbidity. That is, the test is valid if either DAG 3 or DAG 4 generated the data.

Proof: Under this design, the matched sib TDT simply tests whether A is independent of G conditional on C . Thus, the test will be valid if and only if G is d-separated from A given C . As we have seen in example 2, this is the case if and only if true comorbidity is absent.

Propositions 1 and 2 taken together have the following troubling corollary.

Corollary 1: In the presence of both true within-family comorbidity and spurious comorbidity, both the sib TDT test and the matched sib TDT test are invalid. That is, both tests are invalid if DAG 1 generated the data.

THE TDT

The sib TDT test has the advantage that data on parental allelic status are unnecessary but has the disadvantage that unaffected sibs must be available. In contrast, the TDT requires data on parental alleles but does not require the availability of unaffected sibs. We use TDT to refer to any test using cases and their parents to test whether A (the diagnosed target disease) is associated with the candidate gene G . In the simplest version of the TDT, we again find all clinically ascertained subjects with the target disease D and obtain allelic data on the subject and both parents at locus G . To keep the exposition simple, without loss of generality, we suppose that all parental pairs consist of one heterozygous parent (which we denote $G = 1$) and one homozygous parent without the susceptibility allele (which we denote by $G = 0$). Thus, each proband is either homozygous ($G = 0$) or heterozygous ($G = 1$). The TDT then tests whether the number h of heterozygous probands has a binomial ($n, 1/2$) distribution, where n is the total number of probands. We use the same definition of test invalidity as for the sib TDT.

Proposition 3: Suppose assumptions A and B hold. If there exists spurious comorbidity, then the TDT is invalid. If spurious comorbidity is absent, then the TDT is valid even in the presence of true within-family comorbidity.

Proof: By assumption, we restrict attention to the subpopulation of families with one parent heterozygous and one homozygous for the susceptibility allele. By Mendelian laws, G takes the value 1 and 0 with probability 1/2, as all children are either heterozygous ($G = 1$) or homozygous ($G = 0$). Because, by assumption, G is unlinked to D , G and D are independent, so $pr [G = 1 \mid D = 1] = pr [G = 0 \mid D = 1] = 1/2$ under the null hypothesis. However, under the given design, the TDT tests whether $pr [G = 1 \mid A = 1] = 1/2$. Now, $pr [G = 1 \mid A = 1] = pr [G = 1 \mid A = 1, D = 1] pr [D = 1 \mid A = 1] + pr [G = 1 \mid A = 1, D = 0] pr [D = 0 \mid A = 1] = pr [G = 1 \mid A = 1, D = 1]$, where the last equality follows from the fact that we have assumed that there are no false-positive diagnoses, i.e., if $A = 1$, then $D = 1$. Hence, the TDT is valid if and only if G is independent of A given D . However, on DAGs 1–4 G is d-separated from A given D only on the graphs in which the arrow from C to A is absent (i.e., the graphs on which there is no spurious comorbidity).

The next result shows that, unlike the sib TDT test, collecting additional data on the depression status of the probands cannot help. Let TDT_1 be the TDT statistic restricted to probands with the comorbid phenotype and let TDT_0 be the TDT statistic restricted to probands lacking the comorbid phenotype.

Proposition 4: Suppose assumptions A and B hold. Then both TDT_0 and TDT_1 are invalid tests. This is so regardless of whether there is spurious comorbidity, true comorbidity, both, or neither.

Proof: It is clear that if we can show that both tests are invalid when there is neither spurious nor true comorbidity (i.e., DAG 4 generated the data), then it will be invalid in all settings. Now, under the above study design, $pr [G = 1 \mid D = 1] = 1/2$, which implies that $pr [G = 1 \mid A = 1] = 1/2$ because spurious comorbidity is absent. Further, the TDT_1 and TDT_0 , respectively, test the hypothesis that $pr [G = 1 \mid A = 1, C = 1] = 1/2$ and $pr [G = 1 \mid A = 1, C = 0] = 1/2$. Now, by Bayes' rule, $pr [G = 1 \mid A = 1, C = j] = pr [C = j \mid A = 1, G = 1] pr [G = 1 \mid A = 1] / pr [C = j \mid A = 1]$. Hence, for validity of either test, we require that C be independent of G given A . But this is false because on DAG 4, G is a parent of C and this is not d-separated from C given A .

We further note that proposition 4 remains true (by essentially the same proof) even if one has also matched the parents to the probands on the comorbid phenotype.

SUMMARY

The results obtained in this paper are rather sobering. If there exists spurious comorbidity between the target phenotype D and a second phenotype C that itself may be in linkage disequilibrium with the candidate gene G , then both the usual parental and the sib TDT may be invalid tests of linkage disequilibrium between the

gene and the target phenotype. Using simulation, we [Smoller et al., 2000] empirically demonstrated the invalidity of the TDT in the presence of spurious comorbidity and examined the magnitude of the bias in realistic settings as a function of the degree of spurious comorbidity and the strength of the association between the gene G and the comorbid disease C . We showed that the probability of obtaining a misattributed association may sometimes be high. The problem of spurious comorbidity bias may become more acute when, in coming years, whole genome association studies are possible because then it is almost certain that one will be testing genes that would affect a phenotype responsible for spurious comorbidity.

If there is no true within family comorbidity (association) between D and C , a valid sib TDT can be obtained by matching the sib pairs on the comorbid phenotype. However this “fix” may rarely work because the comorbid phenotype that is the cause of the spurious comorbidity may be unknown (and, worse yet, possibly unsuspected), so matching is not possible or may be truly associated with the target phenotype within families.

One way to always eliminate the bias due to spurious comorbidity is to recruit a population-based sample of subjects with the target disease. To obtain a population-based sample, one must take a random sample of the population and test each sampled subject for the target disease. Unfortunately, when population-based estimates are not already available, this design may be prohibitively expensive if the target disease is rare, as a huge sample must then be examined to find a sufficient number of cases. A second method of eliminating the bias is to weight each matched proband-sibling pair in a sib TDT analysis or each proband-parental trio in a TDT analysis by the inverse of the conditional probability $pr(A = 1 \mid D = 1, G)$ of the proband's being ascertained given his/her genotype G . However, little is gained by this approach over the previous method. To see this, note that by Bayes' rule and the assumption of no false-positive diagnoses $pr(A = 1 \mid D = 1, G) = pr(D = 1 \mid A = 1, G) pr(A = 1 \mid G) / pr(D = 1 \mid G) = pr(A = 1 \mid G) / pr(D = 1 \mid G) = pr(G \mid A = 1) pr(A = 1) / pr(G \mid D = 1) pr(D = 1)$ and to obtain precise estimates of $pr(D)$ and $pr(G \mid D = 1)$, we again need to test a large random population sample for the target phenotype D .

If one suspects possible spurious comorbidity attributable to a particular phenotype C , then, in principle, one can test this suspicion by comparing the rate of the comorbid phenotype among those diagnosed with the target phenotype D to the rate in a population sample of subjects with the target phenotype. Unfortunately, such a comparison requires examining a random sample of the population for the target phenotype and so again may be economically unfeasible. Furthermore, it is always possible that there exists spurious comorbidity attributable to an unsuspected and/or unknown phenotype C , in which case the analyst may be totally unaware that parental and sib TDTs may be invalid.

We assumed the diagnostic test for condition D has 100% sensitivity and specificity. Our results are unchanged if such is not the case as long as these do not depend on whether the comorbid phenotype C is present or absent. If sensitivity depends on the comorbid phenotype, this will itself produce ascertainment bias wholly analogous in its effect to the spurious comorbidity studied in this paper.

Finally we assumed that comorbid disease C was not a cause of the target disease D . If this assumption is false and thus there is an arrow from C to D on DAGs 1–4, both nominal α -level parental TDT and sib TDTs will reject at a rate greater than α ,

implying a correlation between the candidate gene G and the target disease D . However, we do not regard this association as artifactual because the candidate gene G is really a contributing cause of (or is in linkage disequilibrium with a contributing cause of) the target disease D because, by assumption, it is a cause of or in linkage disequilibrium with a cause of the comorbid disease C , and C in turn causes D .

ACKNOWLEDGMENTS

Support for this research was provided in part by grants AI32475-09 and MH59532-02 from the National Institutes of Health.

REFERENCES

- Boehnke M, Langefeld CD. 1998. Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *Am J Hum Genet* 62:950–61.
- Cos DR, Wermuth N. 1993. Linear dependencies represented by chain graphs (with discussion). *Stat Sci* 8:204–18, 247–77.
- Curtis D. 1997. Use of siblings as controls in case-control association studies. *Ann Hum Genet* 61:319–33.
- Greenland S, Pearl J, Robins J. 1999. Causal diagrams for epidemiologic research. *Epidemiology* 10:37–48.
- Horvath S, Larid NM. 1998. A discordant sibship test for disequilibrium and linkage: no need for parental data. *Am J Hum Genet* 63:1886–97.
- Lange K, Elston RC. 1975. Extension to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. *Hum Hered* 25:95–105.
- Lauritzen SL, David AP, Larsen BN, Leimar HG. 1990. Independence properties of directed Markov fields. *Networks* 20:491–505.
- Magee W, Eaton W, Wittchen H-U, McGonagle K, Kessler R. 1996. Agoraphobia, simple phobia, and social phobia in the National Comorbidity Survey. *Arch Gen Psychiatry* 53:159–68.
- Pearl J. 1995. Causal diagrams for empirical research. *Biometrika* 82:669–90.
- Pearl J. 2000. *Causality*. Cambridge, England: Cambridge University Press.
- Pearl J, Verma T. 1991. A theory of inferred causation. In: Allen JA, Fikes R, Sandewall E, editors. *Principles of knowledge, representation and reasoning: Proceedings of the Second International Conference*. pp 441–52.
- Robins JM. 1986. A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Math Model* 7:1393–512.
- Robins JM. 2001. Data, design, and background knowledge in etiologic inference. *Epidemiology* 11:313–20.
- Schneier FR, Johnson J, Hornig CD, Liebowitz MR, Weissman MM. 1992. Social phobia: comorbidity and morbidity in an epidemiologic sample. *Arch Gen Psychiatry* 49:282–8.
- Smoller J, Lunetta KL, Robins JM. 2000. Implications of comorbidity and ascertainment bias for identifying disease genes. *Am J Med Genet (Neuropsychiatr Genet)* 96:817–22.
- Spielman RS, Ewens WJ. 1998. A sibship test for linkage in the presence of association: the sib transmission disequilibrium test. *Am J Hum Genet* 62:450–8.
- Spielman R, McGinnis R, Ewens W. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52:506–16.
- Spirtes P, Glymour C, Scheines R. 1993. *Causation, prediction, and search*. New York: Springer-Verlag.
- Terwilliger JD, Ott J. 1992. A haplotype-based “haplotype relative risk” approach to detecting allelic associations. *Hum Hered* 43:337–46.