Hypothesis Tests for Neyman's Bias in Case-Control Studies

D.M. Swanson[1], C.D. Anderson[2], and R.A. Betensky[3]

[1]GNS Healthcare, 196 Broadway, Cambridge, 02139,

[2]Department of Neurology, Massachusetts General Hospital, 55 Fruit St, Boston, Massachusetts, 02114,

[3]Harvard School of Public Health, 655 Huntington Ave, Boston, Massachusetts, 02115.

Abstract

Survival bias is a long-recognized problem in case-control studies, and many varieties of bias can come under this umbrella term. We focus on one of them, termed Neyman's bias or "prevalence-incidence bias." It occurs in case-control studies when exposure affects both disease and disease-induced mortality, and we give a formula for the observed, biased odds ratio under such conditions. We compare our result with previous investigations into this phenomenon and consider models under which this bias may or may not be important. Finally, we propose three hypothesis tests to identify when Neyman's bias may be present in case-control studies. We apply these tests to three data sets, one of stroke mortality, another of brain tumors, and the last of atrial fibrillation, and find some evidence of Neyman's bias in the former two cases, but not the last case.

Keywords: odds ratio; survival bias; truncation.

Introduction.

Survival bias is a frequent source of concern in case-control studies (Sackett, 1979; Rothman et al., 2008). Sackett describes nine types of bias common in case-control studies, and we focus our investigation on one of them, first identified by Jerzy Neyman and now known as Neyman's bias or "prevalence-incidence bias" (Neyman, 1955). It is a bias that occurs when prevalent cases are sampled and exposure affects disease and disease-associated mortality. Since Neyman's article was written in the 1950's when the relationship between smoking and lung cancer was under debate, he uses an example that focuses on that subject. He disregards competing risks and supposes that if, in fact, smoking is protective against lung cancer, but lung cancer mortality is far higher among non-smokers than smokers, then the odds ratio would suggest that smoking is a risk factor for disease as was being observed at the time. In our study of the subject, we focus on three other examples, one coming from a study of brain tumors and chemotherapy, another coming from a GWAS of ischemic stroke, and the last coming from a study of atrial fibrillation in the Framingham Heart Study. Prevalence-incidence bias could arise in the study of brain tumors if certain patients are assessed to have disease too progressed to benefit from chemotherapy and therefore do not undergo treatment. The GWAS could

suffer from prevalence-incidence bias if a certain subset of patients die before admission to a hospital and study entry. We use our study of atrial fibrillation in the Framingham Heart study as an example of a prospective design that should therefore not suffer from prevalence-incidence bias. We consider these data as being generated under the null hypothesis of no prevalence-incidence bias in order to substantiate the validity of the test.

Despite Neyman's early identification of this bias, methodological investigation into it has been limited. Hill (2003) uses a compartment model to show how bias arises when performing case-control studies on prevalent cases if the risk factor impacts both disease and mortality from disease. He also shows that any impact of the risk factor on mortality from other causes does not impact the observed odds ratio, which demonstrates that Neyman was justified in ignoring competing risks. While trying to draw inference on incidence instead of the odds ratio, Fluss et al. (2012), Keiding (1991), and Keiding (2006) all consider the problem of using cross-sectional designs and their resultant sampling biases.

Anderson et al. (2011) performs a computational investigation into Neyman's bias, recognizing that genome-wide association studies (GWAS) and their use of prevalent cases in case-control study designs were susceptible to it. If an allele is a risk factor for both disease and mortality from disease, then the common practice of calculating an odds ratio from prevalent cases and controls could lead to biased inference. Since the odds ratios in such studies are usually small, differences in disease-associated mortality between the exposed and unexposed would not be required for a risk allele to be observed as protective, or vice versa. Their own investigation is motivated by a locus found to be significantly associated with ischemic stroke in longitudinal studies that did not replicate using a case-control design. As a solution, they simulate data under different disease and mortality risk models and then fit regression models for percent bias of the odds ratio to the disease and mortality risk model parameters. These fitted models give researchers a means to investigate the potential biases of estimated odds ratios in their own studies.

In this paper, we propose a framework for consideration of Neyman's bias and examine it from a modeling perspective. We suggest three hypothesis tests to assess whether Neyman's bias is present in a study and then apply these tests to three data sets mentioned: one of brain tumors and chemotherapy, another a GWAS of ischemic stroke, and the last focused on atrial fibrillation in the Framingham Heart Study. We propose hypothesis tests rather than methods to recover the true, unbiased odds ratio, since that quantity is unrecoverable under the framework we consider.

Methods.

2.1 Notation and background.

Suppose that we have a setting similar to that described in Anderson et al. (2011), where we have some binary risk SNP or gene, G, that takes on the value 1 with probability p ("exposed") and 0 with probability $1-p$ ("unexposed"). Let D denote age of disease onset, and suppose that G may be associated with D. Let $\{M_{a,j}\}$, $j = 1, \ldots, n$, denote

age at mortality from all other causes not associated with disease. Let $\{X_i\}$, $i = 1, \ldots, m$, denote latent time from disease onset to the i^{th} mortality cause related to disease and let $X = \min\{X_i\}$. Thus, $M_{a,j} \perp\!\!\!\perp (X \;\; D)^T \mid G$ for all j, where $W \perp\!\!\!\perp Y \mid Z$ denotes statistical independence of W and Y conditional on Z (Dawid, 1979). We define $M_a \equiv \min\{M_{a,j}\}$, and thus $M_a \perp\!\!\!\perp (X \;\; D)^T \mid G$. Let $M_{d,i} \equiv D + X_i$ and $M_d = \min(M_{d,i}) = D + X$. If $X_i \perp\!\!\!\perp D$, then $M_{d,i}$ is necessarily associated with D because $M_{d,i} \equiv D + X_i$ (i.e., $M_{d,i}$ denotes the age at disease-associated mortality cause i). In fact, X_i would have to be associated with D in a specific way to have $M_{d,i} \perp\!\!\!\perp D$. We do not assume X_i is a positive random variable so that we can have $P(M_{d,i} < D) \geq 0$. While it may seem counterintuitive to allow for disease-associated mortality prior to disease, this flexibility fits into a realistic framework. For example, if the disease of interest is stroke, and there exists an association between death from myocardial infarction and stroke, then indeed mortality associated with disease, though not directly caused by it, can occur before disease and can bias the odds ratio, as we show later.

It is not a limitation of this conceptual framework to assume the existences of the $M_{a,j}$'s, ages at causes of mortality unrelated to our disease of interest. They are present to show their lack of effect on the observed odds ratio in the work to follow. Since there is no cap on the possible number of $M_{d,i}$'s, all causes of mortality can be considered as disease-associated if desired by the analyst.

2.2 Formulae.

Suppose we perform a case-control study of prevalent cases at age t^*, and define $C_a \equiv I(t^* \leq M_a)$, $C_d \equiv I(t^* \leq M_d)$, where $I(\cdot)$ is the indicator function, and $C \equiv C_d \times C_a$. While C_d and C_a are functions of M_d and M_a, mortality causes, we can consider M_d and M_a more generally as anything that would render a subject unable to enter the study that is associated and unassociated with disease, respectively. A subject is available to enter the study at age t^* if $C = 1$; i.e., if the subject has not died from any cause by age t^*. Denote the cumulative distribution function associated with random variable Y as $F_Y(t)$. Then the target odds ratio among the population at age t^* is

$$\begin{aligned} OR_{tr}(t^*) &= \frac{P(\text{Case, Exposed})\, P(\text{Control, Unexposed})}{P(\text{Control, Exposed})\, P(\text{Case, Unexposed})} \\ &= \frac{P(D \leq t^*, G = 1) P(D > t^*, G = 0)}{P(D > t^*, G = 1) P(D \leq t^*, G = 0)} \\ &= \frac{F_{D|G=1}(t^*)\, p\, (1 - F_{D|G=0}(t^*))\, (1 - p)}{(1 - F_{D|G=1}(t^*))\, p\, F_{D|G=0}(t^*)\, (1 - p)} \\ &= \frac{F_{D|G=1}(t^*)\, (1 - F_{D|G=0}(t^*))}{(1 - F_{D|G=1}(t^*))\, F_{D|G=0}(t^*)}. \end{aligned}$$

In contrast, the observed odds ratio among prevalent cases at age t^* is

$$\begin{aligned} OR_{ob}(t^*) &= \frac{P(\text{Case, Exposed, Observed})\, P(\text{Control, Unexposed, Observed})}{P(\text{Control, Exposed, Observed})\, P(\text{Case, Unexposed, Observed})} \\ &= \frac{P(D \le t^*, G=1, C=1)\, P(D > t^*, G=0, C=1)}{P(D > t^*, G=1, C=1)\, P(D \le t^*, G=0, C=1)} \\ &= \frac{P(D \le t^*, G=1, C_a=1, C_d=1)\, P(D > t^*, G=0, C_a=1, C_d=1)}{P(D > t^*, G=1, C_a=1, C_d=1)\, P(D \le t^*, G=0, C_a=1, C_d=1)}. \end{aligned}$$

Now consider the term $P(D \le t^*, G=1, C_a=1, C_d=1)$. We can factor the probability as

$$\begin{aligned} &P(D \le t^*, G=1, C_a=1, C_d=1) \\ =&P(D \le t^*, C_d=1 | C_a=1, G=1)\, P(C_a=1|G=1)\, P(G=1). \end{aligned}$$

Since $M_a \perp\!\!\!\perp (X \;\; D)^T \mid G$ and $M_d \equiv X + D$, $M_a \perp\!\!\!\perp M_d \mid G$, and since C_a and C_d are functions of only M_a and M_d (with fixed and known t^*), respectively, $(D \;\; C_d)^T \perp\!\!\!\perp C_a \mid G$. Using this conditional independence, the probability further simplifies to

$$P(D \le t^*, C_d=1|G=1)\, P(C_a=1|G=1)\, P(G=1),$$

which is equal to

$$\int_0^{t^*} (1 - F_{X|D=t,G=1}(t^* - t)) \partial F_{D|G=1}(t)\, (1 - F_{M_a|G=1}(t^*))\, p.$$

Analogous simplifications of the other terms of $OR_{ob}(t^*)$ yield

$$= \frac{\int_0^{t^*} (1 - F_{X|D=t,G=1}(t^* - t)) \partial F_{D|G=1}(t)\, (1 - F_{M_a|G=1}(t^*))\, p}{\int_{t^*}^{\infty} (1 - F_{X|D=t,G=1}(t^* - t)) \partial F_{D|G=1}(t)\, (1 - F_{M_a|G=1}(t^*))\, p} \frac{\int_{t^*}^{\infty} (1 - F_{X|D=t,G=0}(t^* - t)) \partial F_{D|G=0}(t)\, (1 - F_{M_a|G=0}(t^*))\, (1-p)}{\int_0^{t^*} (1 - F_{X|D=t,G=0}(t^* - t)) \partial F_{D|G=0}(t)\, (1 - F_{M_a|G=0}(t^*))\, (1-p)}$$

$$= \frac{\left[\int_0^{t^*} (1 - F_{X|D=t,G=1}(t^* - t)) \partial F_{D|G=1}(t)\right] \left[\int_{t^*}^{\infty} (1 - F_{X|D=t,G=0}(t^* - t)) \partial F_{D|G=0}(t)\right]}{\left[\int_{t^*}^{\infty} (1 - F_{X|D=t,G=1}(t^* - t)) \partial F_{D|G=1}(t)\right] \left[\int_0^{t^*} (1 - F_{X|D=t,G=0}(t^* - t)) \partial F_{D|G=0}(t)\right]}. \quad (1)$$

Note that when $X \perp\!\!\!\perp D \mid G$, we observe

$$\begin{aligned} OR_{ob}(t^*) &= \frac{P(\text{Case, Exposed, Observed})\, P(\text{Control, Unexposed, Observed})}{P(\text{Control, Exposed, Observed})\, P(\text{Case, Unexposed, Observed})} \\ &= \frac{\left[\int_0^{t^*} (1 - F_{X|G=1}(t^* - t)) \partial F_{D|G=1}(t)\right] \left[\int_{t^*}^{\infty} (1 - F_{X|G=0}(t^* - t)) \partial F_{D|G=0}(t)\right]}{\left[\int_{t^*}^{\infty} (1 - F_{X|G=1}(t^* - t)) \partial F_{D|G=1}(t)\right] \left[\int_0^{t^*} (1 - F_{X|G=0}(t^* - t)) \partial F_{D|G=0}(t)\right]}. \end{aligned}$$

This assumption may be reasonable for some exposures that are risk factors for diseases whose course is independent

of the age of onset given G.

Returning to the general case (1), we consider ways in which $OR_{ob}(t^*) = OR_{tr}(t^*)$ holds. Recall that $M_d \equiv D + X$, and that X need not be a positive random variable. Suppose that $X \equiv A - D$, for some positive random variable A independent of D, conditional on G. Then $M_d \equiv D + X = D + (A - D) = A$. So $M_d = A$ and is independent of D given G, or in notation, $M_d \perp\!\!\!\perp D \mid G$ (Dawid, 1979). Notice that when M_d is defined in this way, an association necessarily exists between X and D, conditional on G, since X is itself a function of D. If $M_d \perp\!\!\!\perp D \mid G$ holds, then (1) reduces to

$$\begin{aligned} OR_{ob}(t^*) =& \frac{\left[(1 - F_{M_d|G=1}(t^*)) \int_0^{t^*} \partial F_{D|G=1}(t)\right] \left[(1 - F_{M_d|G=0}(t^*)) \int_{t^*}^{\infty} \partial F_{D|G=0}(t)\right]}{\left[(1 - F_{M_d|G=1}(t^*)) \int_{t^*}^{\infty} \partial F_{D|G=1}(t)\right] \left[(1 - F_{M_d|G=0}(t^*)) \int_0^{t^*} \partial F_{D|G=0}(t)\right]} \\ =& \frac{F_{D|G=1}(t^*)\,(1 - F_{D|G=0}(t^*))}{(1 - F_{D|G=1}(t^*))\,F_{D|G=0}(t^*)} = OR_{tr}(t^*), \end{aligned} \tag{2}$$

where (2) follows from

$$\begin{aligned} F_{X|D=t,G=g}(t^* - t) &= F_{X+t|D=t,G=g}(t^*) \\ &= F_{X+D|D=t,G=g}(t^*) = F_{M_d|D=t,G=g}(t^*) = F_{M_d|G=g}(t^*). \end{aligned}$$

So when $M_d \perp\!\!\!\perp D \mid G$, M_d behaves like M_a in the sense that $OR_{ob}(t^*)$ is no longer a function of the distribution of M_d and $OR_{tr}(t^*) = OR_{ob}(t^*)$. While $M_d \perp\!\!\!\perp D \mid G$ is a sufficient condition for $OR_{tr}(t^*) = OR_{ob}(t^*)$, it is not necessary; there exist multivariate distributions $(X \;\; D \;\; G)^T$ such that $OR_{tr}(t^*) = OR_{ob}(t^*)$, but $M_d \not\perp\!\!\!\perp D \mid G$ (M_d is not independent of D conditional on G). For example, consider the case in which $F_{M_d|D=t,G=g}(x) = 0$ if $x \leq t^*$ or $x \leq t$ and $F_{M_d|D=t,G=g}(x) = 1$, otherwise for $g \in \{0, 1\}$; i.e., no disease-related death occurs prior to t^* and in this way cannot bias $OR_{ob}(t^*)$, but in the region $D > t^*$, M_d is perfectly correlated with D so that $M_d \not\perp\!\!\!\perp D \mid G$. Nonetheless, in § 5 we propose tests of deviations from $M_d \perp\!\!\!\perp D \mid G$ since the cases in which $M_d \not\perp\!\!\!\perp D \mid G$, but $OR_{tr}(t^*) = OR_{ob}(t^*)$ holds are unlikely to occur as is the case in this example.

Scientific hypotheses versus sampling bias hypotheses.

We distinguish between $H_{0S} : OR_{tr}(t^*) = 1$ (at some time t^*, the true odds ratio is one), which we term the "scientific null hypothesis" and $H_{0B} : OR_{tr}(t^*) = OR_{ob}(t^*)$ (there is no bias in the odds ratios at time t^*), which we term the "sampling bias null hypothesis." The alternative hypothesis in each case is the complement of the null hypothesis. We describe characteristics of these hypotheses.

Under $H_{0S} : OR_{tr}(t^*) = 1$ *and* $H_{0S}^c : OR_{tr}(t^*) \neq 1$:

Even if mortality from other causes, M_a, depends on G, it does not affect the bias of the observed odds ratio; in other words, $OR_{ob}(t^*)$ and $OR_{tr}(t^*)$ are not a function of the distribution of M_a. Thus, we may assume, as Neyman (1955) does in his original example and Hill (2003) confirms, that mortality from other causes is not present and death can only occur from disease. Similarly, the probability of exposure, p, does not affect $OR_{ob}(t^*)$. Also, if $F_{M_d|G=g}(t^*) = 0$ for $g \in \{0, 1\}$ (which is the case when no disease-associated mortality occurs prior to t^*), then $OR_{ob}(t^*)$ is unbiased: $OR_{ob}(t^*) = OR_{tr}(t^*)$. This result is expected since it is disease-related mortality that results in the bias-inducing differential selection between the exposed and unexposed.

Under $H_{0S}^c : OR_{tr}(t^*) \neq 1$:

Under the following four conditions, bias exists (i.e., $OR_{ob}(t^*) \neq OR_{tr}(t^*)$):

1. $F_{X|D=t,G=0}(t^* - t) = F_{X|G=0}(t^* - t) = F_{X|G=1}(t^* - t) = F_{X|D=t,G=1}(t^* - t)$ for all t (i.e., the mortality distribution from disease-onset is identical between the exposed and unexposed and not dependent on age at disease-onset).

2. $F_{X|G=g}(t^{**}) > 0$ for some $g \in \{0, 1\}$. In other words, either the exposed or unexposed have positive probability of dying from disease by t^{**}, where t^{**} is defined as the time between t^* and the first possible presence of disease among the exposed or unexposed (i.e., $\inf\{F_{D|G=g}(t) > 0 : t \in [0, \infty), g \in \{0, 1\}\}$) so that the bias-inducing event will have some chance of occurring prior to study entry at age t^*).

3. $P(X > 0) = 1$, implying $P(D < M_d) = 1$.

4. $F_{D|G=0}(x) = F_{D|G=1}(x-k)$ for all x for some $k \neq 0$, and $F_{D|G=0}(t^*) > 0$ or $F_{D|G=1}(t^*) > 0$ (i.e., the disease distributions for the exposed and unexposed are in the same location family, and $k \neq 0$ implies $OR_{tr}(t^*) \neq 1$).

These assumptions seem plausible if some exposure affects the mean age of disease, though the shape of the disease distribution is approximately the same between exposed and unexposed, and after disease occurrence, hazard of mortality is identical among those with and without the exposure and not a function of age at disease onset. The theorem and proof of this result is found in the Appendix (Theorem 1). Additionally, in that proof we examine the direction of bias; we find that when $OR_{tr}(t^*) < 1$, then $OR_{ob}(t^*) > OR_{tr}(t^*)$, and when $OR_{tr}(t^*) > 1$, then $OR_{ob}(t^*) < OR_{tr}(t^*)$. Thus, if the degree of bias is relatively small, then it can be viewed as a bias toward an observed odds ratio of 1. However, $OR_{ob}(t^*)$ is by no means bounded by 1 and so if the amount of bias is great, $OR_{ob}(t^*)$ and $OR_{tr}(t^*)$ can lie on opposite sides of 1, leading to wrongly inferring a truly protective exposure as a risk factor for the outcome or a true risk factor as protective against the outcome.

This result of $OR_{ob}(t^*) \neq OR_{tr}(t^*)$ will not necessarily hold if conditions $1 - 3$ hold, but condition 4 is not satisfied (the distributions of disease of exposed and unexposed are not in the same location family). Under such a scenario, there may not be bias, as Example 1 in the Appendix illustrates. Additionally, if we only assume that conditions $2 - 3$ are satisfied, then there may or may not be bias. See Examples 2 and 3 in the Appendix for instances

of $OR_{ob}(t^*) = OR_{tr}(t^*)$ and $OR_{ob}(t^*) \neq OR_{tr}(t^*)$, respectively, when X is associated with G (but is independent of D given G: $X \perp\!\!\!\perp D \mid G$). It follows that if there exist no conditional independences, one can make no conclusions regarding the relationship between $OR_{tr}(t^*)$ and $OR_{ob}(t^*)$ as there is even greater flexibility in the joint model. Lastly, if only $X \perp\!\!\!\perp G \mid D$ is assumed so that X may depend on D (i.e., time to disease-induced mortality may depend on age at disease-onset), again $OR_{tr}(t^*)$ and $OR_{ob}(t^*)$ may or may not be equal. This result follows from the proof with location families and Example 1 because they are special cases of only assuming $X \perp\!\!\!\perp G \mid D$.

Under $H_{0S} : OR_{tr}(t^*) = 1$:

If we only assume that $OR_{tr}(t^*) = 1$ with no conditions on $OR_{tr}(t)$ for $t < t^*$, and also that $X \perp\!\!\!\perp D \mid G$ and $F_{X|G=0}(t) \neq F_{X|G=1}(t)$ for some $t < t^{**}$, one cannot conclude anything regarding the relationship between $OR_{tr}(t^*)$ and $OR_{ob}(t^*)$. Consider Examples 4 and 5 in the Appendix for instances of $OR_{ob}(t^*) = OR_{tr}(t^*) = 1$ and $OR_{ob}(t^*) \neq OR_{tr}(t^*) = 1$, respectively. We also observe that if $OR_{tr}(t) = 1$ for all $t \leq t^*$ and $F_{X|D,G=0}(t) = F_{X|D,G=1}(t)$ for all $t < t^{**}$, $OR_{tr}(t^*) = OR_{ob}(t^*) = 1$.

The odds ratio when T^* is not fixed.

If the case-control study consists of people of many ages, then t^*, previously considered fixed, can be considered random. Let us denote this random variable T^*. Under these conditions, the target odds ratio becomes

$$\begin{aligned} OR_{tr}(T^*) &= \frac{P(\text{Case, Exposed})\, P(\text{Control, Unexposed})}{P(\text{Control, Exposed})\, P(\text{Case, Unexposed})} \\ &= \frac{P(D \leq T^*, G = 1)\, P(D > T^*, G = 0)}{P(D > T^*, G = 1)\, P(D \leq T^*, G = 0)} \\ &= \frac{\int F_{D|G=1}(u)\, \partial F_{T^*}(u)\, p\, (1 - \int F_{D|G=0}(u)\, \partial F_{T^*}(u))\, (1-p)}{(1 - \int F_{D|G=1}(u)\, \partial F_{T^*}(u))\, p \int F_{D|G=0}(u)\, \partial F_{T^*}(u)\, (1-p)} \\ &= \frac{\int F_{D|G=1}(u)\, \partial F_{T^*}(u)\, (1 - \int F_{D|G=0}(u)\, \partial F_{T^*}(u))}{(1 - \int F_{D|G=1}(u)\, \partial F_{T^*}(u))\, \int F_{D|G=0}(u)\, \partial F_{T^*}(u)}. \end{aligned}$$

Making no assumptions about the joint model $(D \;\; X \;\; M_a \;\; G)^T$, the observed odds ratio is

$$\begin{aligned} OR_{ob}(T^*) &= \frac{P(\text{Case, Exposed, Observed})\, P(\text{Control, Unexposed, Observed})}{P(\text{Control, Exposed, Observed})\, P(\text{Case, Unexposed, Observed})} \\ &= \frac{P(D \leq T^*, T^* < M_d, T^* < M_a, G = 1)\, P(D > T^*, T^* < M_d, T^* < M_a, G = 0)}{P(D > T^*, T^* < M_d, T^* < M_a, G = 1)\, P(D \leq T^*, T^* < M_d, T^* < M_a, G = 0)} \end{aligned}$$

$$= \frac{P(D \leq T^*, T^* < M_d, T^* < M_a | G = 1)\, P(G = 1)}{P(D > T^*, T^* < M_d, T^* < M_a | G = 1)\, P(G = 1)}$$

$$\frac{P(D > T^*, T^* < M_d, T^* < M_a | G = 0)\, P(G = 0)}{P(D \leq T^*, T^* < M_d, T^* < M_a | G = 0)\, P(G = 0)}$$

$$= \frac{P(D \leq T^*, T^* < M_d, T^* < M_a | G = 1)\, P(D > T^*, T^* < M_d, T^* < M_a | G = 0)}{P(D > T^*, T^* < M_d, T^* < M_a | G = 1)\, P(D \leq T^*, T^* < M_d, T^* < M_a | G = 0)}.$$

While $P(G = 1) = p$ cancels from $OR_{ob}(T^*)$ as before with $OR_{ob}(t^*)$, we see that even if $(D \;\; X)^T \perp\!\!\!\perp M_a \mid G$, we cannot factor $P(T^* < M_a | G = g)$ out of the expression. So $OR_{ob}(T^*)$ becomes a function of M_a, causes of mortality unassociated with the disease under investigation. Additionally, regardless of whether $P(T^* < M_a | G = g)$ factors out of the expression, $D \perp\!\!\!\perp M_d \mid G$, which we have stated before as being sufficient for $OR_{ob}(t^*) = OR_{tr}(t^*)$, is not sufficient for $OR_{ob}(T^*) = OR_{tr}(T^*)$. This point is relevant as we propose hypothesis tests below. This is also important to remember since $OR_{ob}(T^*)$ is generally what would be measured in a real-world case-control study where ages of subjects vary. While outside the scope of this investigation, investigators might want to only pool those groups of subjects where $F_{D|G}(\cdot)$ is relatively constant across their age ranges, in which case $D \perp\!\!\!\perp M_d \mid G$ would be sufficient for $OR_{ob}(T^*) = OR_{tr}(T^*)$. Also, if the sample size of a case-control study is sufficient, stratifying subjects by age and calculating age-specific odds ratios would be another way to be assured that $D \perp\!\!\!\perp M_d \mid G$ is sufficient for those stratum-specific odds ratios.

Hypothesis testing.

5.1 Description.

We develop three methods for testing for the presence of Neyman's bias in a study. Again, the "bias null hypothesis" of these tests is $OR_{tr}(t^*) = OR_{ob}(t^*)$, and the alternative is $OR_{tr}(t^*) \neq OR_{ob}(t^*)$. While power may vary as a function of $OR_{tr}(t^*)$, the tests we propose are valid under all values of $OR_{tr}(t^*)$. Each of these three methods makes use of characteristics unique to the data when Neyman's bias is absent, and each test may be more fitting to use than the other two under certain study designs. So, for example, Tests 1 and 2 require study observations to have some variation in age at study entry, a random variable we denote T^*, while Test 3 does not, though Test 3 requires external knowledge of population prevalence of disease and exposure, while neither Test 1 nor Test 2 does.

We have demonstrated above that $M_d \perp\!\!\!\perp D \mid G$ is a sufficient condition for $OR_{tr}(t^*) = OR_{ob}(t^*)$. Ideally, we would have data on all of D, M_d, and G and could test for conditional independences. However, in practice, it may be unlikely that one would have follow-up data on controls and perhaps even cases, in which case M_d would be unknown for one or both groups. Thus, we propose these tests with real-world data limitations in mind.

The first two hypothesis tests we propose attempt to test whether this independence condition holds. Both of these hypothesis tests make use of previous work coming from the truncation methodology literature for tests of "quasi-independence," which refers to independence of random variables in a certain "observable" region of their joint distribution, which we explain further below (Martin and Betensky, 2005; Tsai, 1990). Tests for quasi-independence are based on U-statistics, a class of statistics with broad application outside of these tests and first described in Hoeffding

et al. (1948).

The last hypothesis test we propose assumes $P(D < M_d) = 1$, which may be unreasonable in some settings, but reasonable in others, and depends on whether causes of mortality associated with disease can come before disease onset. The test uses the fact that with data collected under a case-control study design along with population disease prevalence, one can estimate the population exposure proportion. If one has knowledge of the true exposure proportion, any comparison between the true, known value and the calculated quantity can reveal bias in the odds ratio from which it was calculated. Thus, in contrast to the first two tests that detect a sufficient, though not necessary, condition for $OR_{tr}(t^*) = OR_{ob}(t^*)$, rendering the test potentially slightly conservative (though likely not very conservative), this last test achieves its nominal type 1 error rate under the null of $OR_{tr}(t^*) = OR_{ob}(t^*)$ and has power greater than it under the alternative of $OR_{tr}(t^*) \neq OR_{ob}(t^*)$.

5.2 Test 1: Testing for "quasi-independence" using D and M_d.

We are interested in testing independence of D and M_d given G, and our observable region is $D < T^* < M_d$ given G; i.e., realizations of observed (because $T^* < M_d$) cases (because $D < T^*$) of a given exposure status. To accomplish this, we modify a U-statistic test of association of Austin et al. (2013), whose null hypothesis assumes in our context mutual independence of D, T^*, M_d; this is stronger than our null hypothesis. This is a valid approach to testing $D \perp\!\!\!\perp M_d \mid G$, which is sufficient for no Neyman's bias, because $D \perp\!\!\!\perp M_d \mid G$ necessarily implies independence in the region we are defining as observable, $D < T^* < M_d$ given G. Additionally, we focus on the cases in the study, under the assumption that follow-up data on M_d is more likely to be available among them. While the power of this test may suffer in comparison to one that makes use of all observations, the approach makes fewer assumptions on data availability, and in settings in which $P(D < M_d)$ is close to 1, power will not suffer significantly.

To implement the hypothesis test, first we categorize all causes of mortality as M_d, since if D and M_d are associated given G, and $D \perp\!\!\!\perp M_a \mid G$, then categorizing M_a as M_d will maintain that association and avoid the need to censor observations. Doing so is not an approximation nor does it invalidate the test; rather, the test could become invalid if mortality related to disease (M_d) are incorrectly categorized as unrelated to disease (M_a). Also, if $D \perp\!\!\!\perp M_d \mid G$ and $D \perp\!\!\!\perp M_a \mid G$, categorizing M_a as M_d will maintain $D \perp\!\!\!\perp M_d \mid G$. This approach is also legitimate from the perspective that M_d was originally defined as causes of mortality potentially, though not necessarily, associated with disease. Now suppose that we have $1, \ldots, n$ realizations of $(G_i \; D_i \; T_i^* \; M_{d,i})^T$, all cases so that one can assume $D < T^*$ and on whom there is follow-up so M_d is known, and that $C_{ij}^0 = 1$ (alternatively, $C_{ij}^1 = 1$) if $G = 0$ (alternatively, $G = 1$) and $\max\{D_i, D_j\} \leq \min\{M_{d,i}, M_{d,j}\}$, the *comparability* criterion, is satisfied, and $C_{ij}^0 = 0$ (alternatively, $C_{ij}^1 = 0$) otherwise. Define $n_0 \equiv \sum_{i=1}^{n-1} \sum_{j=(i+1)}^{n} C_{ij}^0$ and $n_1 \equiv \sum_{i=1}^{n-1} \sum_{j=(i+1)}^{n} C_{ij}^1$.

The test statistic for the stratum $G = g$, T_g, with $g \in \{0, 1\}$, is

$$T_g = \frac{1}{n_g} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{sgn}((D_i - D_j)(M_{d,i} - M_{d,j}))\, C_{ij}^g.$$

Then $T_g \sim N(0, v_g)$, where

$$v_g = E(\text{sgn}((D_1 - D_2)(M_{d,1} - M_{d,2})\,(D_1 - D_3)(M_{d,1} - M_{d,3}))\, C_{12}^g\, C_{13}^g \,|\, G = g) - (\tau_{D,M_d}^g \mu_{D,M_d})^2$$

and where $\tau_{D,M_d}^g = E(\text{sgn}((D_1 - D_2)(M_{d,1} - M_{d,2})) \mid C_{12} = 1, G = g)$ and $\mu_{D,M_d} = P(C_{12} = 1)$, with $\text{sgn}(x) = 1$ for $x > 0$, -1 for $x < 0$, and 0 for $x = 0$.

Since we would reject if either T_0 or T_1 falls in some predetermined critical region because dependence between D and M_d given either $G = 0$ or $G = 1$ may mean $OR_{tr}(t^*) \neq OR_{ob}(t^*)$, in order to achieve a size α test, we can use a p-value threshold of α^* for T_0 and T_1, where α^* satisfies the equation $\alpha = 1 - (1 - \alpha^*)^2$. So we propose a test that rejects for $\max\{\text{abs}(T_0/v_0^{1/2}), \text{abs}(T_1/v_1^{1/2})) > z_{1-\alpha^*/2}$, where $\text{abs}(x)$ denotes the absolute value of x and $z_{1-\alpha^*/2}$ is the $(z_{1-\alpha^*/2})^{th}$ quantile of a standard normal random variable.

Also, though $D \perp\!\!\!\perp M_d \mid G$ characterizes a subset of situations for which $OR_{tr}(t^*) = OR_{ob}(t^*)$, our test is likely not overly conservative. The majority of situations in which $OR_{tr}(t^*) = OR_{ob}(t^*)$ holds result from $D \perp\!\!\!\perp M_d \mid G$ being satisfied.

Power curves for Test 1 as a function of the association between D and M_d are shown in Figs. 1 and 2. These curves were generated at 11 different values of Kendall's τ, used as a measure of the association between D and M_d. In our case a Kendall's τ value of 0 corresponds to independence between D and M_d, and the power of the test at that value demonstrates the desired type 1 error rate of 0.05. The power curve in Fig. 1 was generated using 3000 iterations at each value, while that in Fig. 2 was generated using 1000 iterations at each value, and power was estimated by averaging over these iterations. In Fig. 1, at each iteration the test statistic was calculated using 1000 comparable pairs (i.e., those pairs that satisfy the comparability criterion mentioned in the description of Test 1). In Fig. 2, the test statistic was calculated using approximately 670 comparable pairs at each iteration–a subset of the 1000 comparable pairs used for Test 2, described below, that satisfied Test 1's more stringent comparability criterion. Fig. 2 also demonstrates a type 1 error rate of 0.05. D, T, and M_d were all distributed normal with means of 5, 9, and 9, respectively, and standard deviations of 0.7.

We now consider realistic settings in which age of entry, T^*, is random. In § 4 above, we saw that the distribution for M_a did not factor out of the odds ratio even when $(D \quad X) \perp\!\!\!\perp M_a \mid G$, and that additionally even under the assumption of $D \perp\!\!\!\perp M_d \mid G$, whether or not the previous assumption held, $OR_{ob}(T^*)$ could be biased; we needed a fixed t^* for these conditional independencies to result in $OR_{tr}(t^*) = OR_{ob}(t^*)$. Thus, it may seem illogical to propose a test that requires variation in T^*, which is precisely when the odds ratio will almost certainly be biased as

shown in § 4. If we do find that $D \perp\!\!\!\perp M_d \mid G$, sufficient for no Neyman's bias, we would need to then stratify our sample according to similar values of T^* such that, within each stratum, T^* can be effectively considered fixed, and then calculate the odds ratio for these different strata. We could then combine these strata into a average odds ratio if desirable or just consider each stratum-specific odds ratio separately.

5.3 Test 2: Testing "quasi-independence" with D and T^*.

We now describe a test related to Test 1, which does not require knowledge of M_d and again focuses on cases, those observations for whom $D < T^*$ is true. Such a test is appropriate if a study did not obtain follow-up on subjects, but did record age at onset of disease for cases. The foundation for the test is based on causal directed acyclic graphs (DAGs), borrowed from the causal inference literature (Hernan and Robins, ress). We use DAGs not for the sake of justifying causal interpretations of OR_{ob}, but rather as a convenient means of encoding conditional independencies. If DAGs are unfamiliar with the reader, Hernan and Robins (ress) describes them well.

By definition, the event $I(T^* < M_d) = 1$ must be satisfied for any subject in the study and can therefore be treated as a conditioning event. Additionally, by definition of $I(T^* < M_d)$, there exists an association between it and both T^* and M_d; to borrow language from the causal inference literature, $I(T^* < M_d)$ is called a "collider" in this instance because both T^* and M_d cause it. Thus, we see in Figs. 3 and 4 arrows between these random variables, indicative of a possible association, and a square around $I(T^* < M_d)$, indicative of a conditioning event. Assuming $0 < P(T^* < M_d) < 1$ so the conditioning event is non-trivial, an association between D and T^* given G implies $D \not\perp\!\!\!\perp M_d \mid G$, and the converse of this statement is also true. Association between D and T^* given G therefore serves as a powerful and valid proxy for association between D and M_d given G. These associations result from conditioning on the "collider" $I(T^* < M_d)$; association paths are opened between D and T^* given G. Were $I(T^* < M_d)$ not to be conditioned on, D and T^* would be independent given G. The structure of this DAG is identical to that found in classic selection bias (even if the variables are not), where exposure and outcome both cause some indicator that is conditioned on in the analysis, which results in a spurious association between exposure and outcome even under the null.

We could assume D and T^* are known for all observations in our data set and propose a test of association for these random variables under the framework described above. However, doing so is unrealistic as it assumes follow-up data on age at disease, D, for those observed at T^* as controls (i.e., those with $D > T^*$). Thus, we assume D and T^* are observed only for cases (i.e., those realizations satisfying $D < T^*$) and propose a test of "quasi-independence" between D and T^* given G in the region of $D < T^*$ given G. If we assume that the independence which holds on the region $D < T^*$ also holds for the entire joint distribution of $(D \;\; T^*)^T$, then since there is an association between D and T^* given G if and only if D and M_d are associated given G, this test is valid. We do not feel this assumption is an overly strong one, but is instead reasonable. In general, a joint distribution that exhibits dependence will not have that structure isolated to a certain region–examination of just a single region will reveal it, even if there are pathological

counterexamples where this behavior does not hold.

We describe here this proposed test of quasi-independence. As before, let there be n realizations of $(G_i\ D_i\ T_i^*)^T$, and again define B_{ij}^0 (alternatively, B_{ij}^1) similarly to how we did with C_i^0 (alternatively, C_i^1), where $B_{ij}^0 = 1$ if $G = 0$ and $\max\{D_i, D_j\} \leq \min\{T_i^*, T_j^*\}$ and $B_{ij}^0 = 0$ otherwise, and where $B_{ij}^1 = 1$ if $G = 1$ and $\max\{D_i, D_j\} \leq \min\{T_i^*, T_j^*\}$ and $B_{ij}^1 = 0$ otherwise. Also, define $m_0 \equiv \sum_{i=1}^{n-1} \sum_{j=(i+1)}^{n} B_{ij}^0$ and $m_1 \equiv \sum_{i=1}^{n-1} \sum_{j=(i+1)}^{n} B_{ij}^1$. Then the test statistic for the stratum $G = g$, W_g, with $g \in \{0, 1\}$, is

$$W_g = \frac{1}{m_g} \sum_{i=1}^{n-1} \sum_{j=(i+1)}^{n} \text{sgn}((D_i - D_j)(T_i^* - T_j^*))\, B_{ij}^g.$$

Then $W_g \sim N(0, u_g)$, where

$$u_g = E(\text{sgn}((D_1 - D_2)(T_1^* - T_2^*)\,(D_1 - D_3)(T_1^* - T_3^*))B_{12}^g\, B_{13}^g \,|\, G = g) - (\tau_{D,T^*}^g \mu_{D,T^*})^2$$

and where $\tau_{D,T^*}^g = E(\text{sgn}((D_1 - D_2)(T_1^* - T_2^*)) \mid B_{12} = 1, G = g)$ and $\mu_{D,T^*} = P(B_{12} = 1)$. As with Test 1, since we would reject if either W_0 or W_1 falls in some predetermined critical region because dependence between D and M_d given either $G = 0$ or $G = 1$ may mean $OR_{tr}(t^*) \neq OR_{ob}(t^*)$, for a size α test, our p-value threshold α^* for W_0 and W_1 satisfies $\alpha = 1 - (1 - \alpha^*)^2$. Thus, our test rejects for $\max\{\text{abs}(W_0/u_0^{1/2}),\ \text{abs}(W_1/u_1^{1/2})) > z_{1-\alpha^*/2}$.

Power curves for Test 2 as a function of the association between D and M_d are shown in Figs. 1 and 2. As with the simulations for Test 1, these curves were generated at 11 different values of Kendall's τ, used as a measure of the association between D and M_d. In our case a Kendall's τ value of 0 corresponds to independence between D and M_d. The power curve for Test 2 in Fig. 1 was generated using 3000 iterations at each value, while that in Fig. 2 was generated using 1000 iterations at each value, and power was again estimated by averaging over these iterations. In both Figs. 1 and 2, at each iteration the test statistic was calculated using 1000 comparable pairs. D, T, and M_d were distributed multivariate normal with means of 5, 9, and 9, respectively, and standard deviations of 0.7. The correlation between D and M_d varied as measured by Kendall's τ, while T was assumed independent of $(T,\ \ M_d)^T$.

As mentioned at the end of the description of Test 1 and for reasons given there, if this test does not reject $D \perp\!\!\!\perp T^* \mid G$, implying $D \not\perp\!\!\!\perp M_d \mid G$, we would again need to stratify the data by T^* in order for $OR_{ob}(t^*)$ to be unbiased for $OR_{tr}(t^*)$. In other words, one could split the data in subsets based on age at entry, T^*, and calculate stratum-specific odds ratios. It is important to note that the reason for doing so applies no more in the context of testing for Neyman's bias than it would any standard case-control study, where a mixture of ages at study entry results in an odds ratio that is difficult to interpret and possibly biased as shown in § 4.

5.4 Test 3: Estimating population exposure proportion.

With knowledge of disease prevalence, we can construct an estimate of the exposure in the general population from case-control study data that is unbiased in the absence of Neyman's bias, but biased otherwise. Thus, if the exposure proportion in the population is also known, as might be the case in GWAS where minor allele frequencies (MAFs) are oftentimes known for SNPs in different populations, we can test for the presence of Neyman's bias by examining their discrepancy. We develop one possible hypothesis test below where, again, H_0 is $OR_{tr}(t^*) = OR_{ob}(t^*)$, and H_a is the complement of H_0.

If we make an assumption of $P(D < M_d) = 1$, then in comparing $OR_{tr}(t^*)$ and $OR_{ob}(t^*)$, we see that their equivalence depends on

$$\frac{F_{D|G=1}(t^*)\,(1 - F_{D|G=0}(t^*))}{(1 - F_{D|G=1}(t^*))\,F_{D|G=0}(t^*)} = \frac{\left[\int_0^{t^*}(1 - F_{X|D=t,G=1}(t^* - t))\partial F_{D|G=1}(t)\right]\left[(1 - F_{D|G=0}(t^*))\right]}{\left[(1 - F_{D|G=1}(t^*))\right]\left[\int_0^{t^*}(1 - F_{X|D=t,G=0}(t^* - t))\partial F_{D|G=0}(t)\right]} \tag{3}$$

if and only if

$$\frac{F_{D|G=1}(t^*)}{F_{D|G=0}(t^*)} = \frac{\int_0^{t^*}(1 - F_{X|D=t,G=1}(t^* - t))\partial F_{D|G=1}(t)}{\int_0^{t^*}(1 - F_{X|D=t,G=0}(t^* - t))\partial F_{D|G=0}(t)}$$

So define $p_2(t^*) \equiv P(G = 1|D < t^*) = P(G = 1 \mid \text{Case at } t^*)$. Then defining

$$h(t^*) \equiv \frac{F_{D|G=1}(t^*)}{F_{D|G=0}(t^*)} = \frac{P(G = 1 \mid \text{Case at } t^*)}{P(G = 0 \mid \text{Case at } t^*)} = \frac{P(G = 1 \mid \text{Case at } t^*)}{1 - P(G = 1 \mid \text{Case at } t^*)},$$

we have $p_2(t^*) = h(t^*)/(1 + h(t^*))$, and defining

$$h^*(t^*) \equiv \frac{\int_0^{t^*}(1 - F_{X|D=t,G=1}(t^* - t))\partial F_{D|G=1}(t)}{\int_0^{t^*}(1 - F_{X|D=t,G=0}(t^* - t))\partial F_{D|G=0}(t)} = \frac{P(G = 1 \mid \text{Case at } t^*, \text{Not censored from disease by } t^*)}{P(G = 0 \mid \text{Case at } t^*, \text{Not censored from disease by } t^*)},$$

then we have $p_2^N(t^*) \equiv P(G = 1 \mid \text{Case at } t^*, \text{Not censored from disease by } t^*) = h^*(t^*)/(1 + h^*(t^*))$. When equation 3 does not hold,

$$p_2^N(t^*) \equiv \frac{h^*(t^*)}{1 + h^*(t^*)} \neq \frac{h(t^*)}{1 + h(t^*)} \equiv p_2(t^*).$$

Thus, if bias is present so that $OR_{tr}(t^*) \neq OR_{ob}(t^*)$, then $h(t^*) \neq h^*(t^*)$, and it will follow that $p_2(t^*) \neq p_2^N(t^*)$. This idea can be leveraged in a hypothesis test if there is external knowledge of the population exposure proportion and population prevalence of disease.

By definition of $p_2^N(t^*)$, its estimator, $\widehat{p_2}^N(t^*)$, is the observed exposure proportion among cases where $E[\widehat{p_2}^N(t^*)] =$

$p_2^N(t^*)$. Let $p_1(t^*) \equiv P(G = 1 \mid D > t^*, M_d > t^*)$, and since $P(D < M_d) = 1$ by assumption, $p_1(t^*) = P(G = 1 \mid D > t^*) = P(G = 1 \mid \text{Control at } t^*)$. Then $\widehat{p}_1(t^*)$ is the observed exposure proportion among controls, and $E[\widehat{p}_1(t^*)] = p_1(t^*)$.

We will estimate $p^N(t^*) \equiv p_1(t^*)\,(1 - p^*(t^*)) + p_2^N(t^*)\,p^*(t^*)$ with $\widehat{p}_1(t^*)\,(1 - p^*(t^*)) + \widehat{p}_2^N(t^*)\,p^*(t^*)$. Also, define $p^*(t^*) \equiv P(\text{Case at } t^*) = P(D < t^*)$, which implies $(1 - p^*(t^*)) = P(\text{Control at } t^*) = P(D > t^*)$. So $p^*(t^*)$ is the population prevalence of disease at a common age t^* and is considered fixed and known. Since

$$\begin{aligned} P(G = 1) &= P(G = 1 \mid D > t^*)\,P(D > t^*) + P(G = 1 | D < t^*)\,P(D < t^*) \\ &= p_1(t^*)\,(1 - p^*(t^*)) + p_2(t^*)\,p^*(t^*), \end{aligned}$$

if $p_2(t^*) = p_2^N(t^*)$, which indicates that $OR_{ob}(t^*) = OR_{tr}(t^*)$, then $p^N(t^*) = P(G = 1)$. Since we consider $P(G = 1)$ fixed and known, the discrepancy between $\widehat{p}^N(t^*)$ and $P(G = 1)$ will inform our test.

Define $\delta(t^*) = p_2(t^*) - p_2^N(t^*)$. Then

$$\begin{aligned} p^N(t^*) + \delta(t^*)\,p^*(t^*) =& p_1(t^*)\,(1 - p^*(t^*)) + p_2^N(t^*)\,p^*(t^*) + \delta(t^*)\,p^*(t^*) \\ =& p_1(t^*)\,(1 - p^*(t^*)) + p_2^N(t^*)\,p^*(t^*) + (p_2(t^*) - p_2^N(t^*))\,p^*(t^*) \\ =& p_1(t^*)\,(1 - p^*(t^*)) + p_2(t^*)\,p^*(t^*) = P(G = 1). \end{aligned}$$

So $P(G = 1)$ and $p^N(t^*)$ differ by $\delta(t^*)\,p^*(t^*)$. The variance associated with our estimate of the exposure proportion $\widehat{p}^N(t^*)$ is

$$v \equiv Var(\widehat{p}^N(t^*)) = (p^*(t^*))^2\left[\frac{p_2^N(t^*)(1 - p_2^N(t^*))}{n_2}\right] + (1 - p^*(t^*))^2\left[\frac{p_1(t^*)(1 - p_1(t^*))}{n_1}\right],$$

where n_2 is the number of cases and n_1 the number of controls. We can estimate v with $\hat{p}_1(t^*)$ and $\widehat{p}_2^N(t^*)$ and call the quantity $\hat{v}$. So, using a large sample approximation, we can construct an α level hypothesis test for the presence of Neyman's bias by rejecting for

$$\left|\frac{P(G = 1) - \widehat{p}^N(t^*)}{\hat{v}^{1/2}}\right| \sim |Z| > z_{1-\alpha/2}.$$

The power becomes

$$\left|\frac{P(G = 1) - \widehat{p}^N(t^*)}{\hat{v}^{1/2}}\right| > z_{1-\alpha/2}$$

$$\approx \frac{P(G = 1) - (\widehat{p}^N(t^*) + \delta(t^*)p^*(t^*))}{\hat{v}^{1/2}} \sim Z > (z_{1-\alpha/2} - \delta(t^*)p^*(t^*))/\hat{v}^{1/2}),$$

assuming one tail probability negligible. We see that power decreases as $p^*(t^*)$ decreases and increases with $\delta(t^*)$, interpreted as the "degree of Neyman's bias."

Power curves for Test 3 are shown in Fig. 5. Consistent with our understanding of the test, power increases as $p^*(t^*)$ increases. The type 1 error rate of the test is 0.05. The x-axis of Fig. 5, relative probability of observation, is defined as $(1 - F_{M_d|G=0}(t^*))/(1 - F_{M_d|G=1}(t^*))$; the x-axis begins at one and values greater than one imply that exposed cases are less likely to be sampled than unexposed cases because exposed subjects get disease earlier. Curves were generated using 300 cases and 300 controls in each simulated study, and the population-level exposure proportion was 0.09. Power was calculated based on 4000 iterations at each of the 11 total relative probabilities seen on the x-axis. This entire procedure was done for disease prevalences, $p^*(t^*)$, of 0.1, 0.2, and 0.3. We assume that there is no variation in t^* so that $p^*(t^*)$ is also fixed.

Data analysis.

6.1 Test 1 applied to an atrial fibrillation data set.

We applied Test 1 to the Framingham Heart study, a longitudinal study where we would not expect to find evidence of Neyman's bias; subjects would not be lost to high-mortality diseases by nature of the study design. We would therefore hope that the null hypothesis for Test 1 would not be rejected for each exposure stratum. We consider our exposure to be gender and the disease atrial fibrillation. Among the cohort, there were 82 males and 188 females who had had atrial fibrillation and died of a related cause. Z-statistics calculated using the methodology of Test 1 for the male and female strata are 0.84 and 0.68, respectively, with associated p-values of 0.20 and 0.25. The interpretation of this result is that, were we to calculate an odds ratio for atrial fibrillation at some set age where the exposure is gender, we would not expect to encounter bias. Again, however, the test is used solely for illustration in this case because the study is prospective and not subject to the bias.

6.2 Test 2 applied to a brain tumor data set.

We apply Test 2 to a brain tumor data set. Seventy-five subjects with oligodendroglioma, a malignant brain tumor, were enrolled in a study at the London Regional Cancer Centre from 1984-1999 (Betensky et al., 2003; Ino et al., 2001). The data set consisted of patient age at diagnosis of oligodendroglioma (i.e., age at disease, D) and age at start of chemotherapy (i.e., entry into the study, T^*) in addition to genetic markers and other covariates. We consider the marker at the 1pLOH locus, thought to be potentially associated with tumor sensitivity to chemotherapy. Applying Test 2 to the data set, first within the exposed stratum of the 1pLOH marker, we obtain a Z-statistic of 6.85, significant at the 0.05 level ($p < 0.001$). The sample size was insufficient to apply the test to the unexposed stratum. However, since a significant test statistic within any stratum is sufficient for rejection of the null hypothesis, we reject the null

hypothesis of $D \perp\!\!\!\perp M_d \mid G$ and conclude that there could be an association between D and M_d within strata of G. The result of the test suggests that if one were to calculate an odds ratio of oligodendroglioma for the 1pLOH marker for subjects at a fixed age, the result might be biased.

Consistent with the comparability criterion of Test 1 being more strict than that for Test 2, the sample size of 75 subjects was insufficient to additionally use Test 1. Had it been used, conclusions drawn from it would not have changed a lack of faith put on the odds ratio in this data set due to results from Test 2.

6.3 Test 3 applied to a stroke-mortality data set.

We apply Test 3 to a GWAS data set of ischemic stroke coming from a cohort based at Massachusetts General Hospital consisting of 383 cases and 384 controls. We use a wide interval estimate of ischemic stroke prevalence, ranging from 0.5%-5%, based on a search of the stroke literature (Feigin et al., 2009; Johnston et al., 2009; CDC, 2012). With this range of $p^*(t^*)$, we reconstruct what would be population exposure proportion, which is unbiased for the true population exposure proportion assuming that Neyman's bias is not present. We calculate a test statistic based on the difference between the true population exposure proportion and our estimate of it, divided by an estimate of the standard error. Using a 0.0005 Bonferroni-adjusted significance level, we find that 42 of the 99 SNPs in the study suggest that Neyman's bias may be present. The interpretation of this result is that any one of the odds ratios calculated for these 42 SNPs might be biased. We additionally perform a power calculation for this test using realized minor allele frequencies in the data set and generous estimates of both $\hat{\nu}$, $\delta(t^*)$, and $p^*(t^*)$. Doing so yields power calculations little above α, at 0.06, which we discuss below.

Discussion.

Test 2 with the brain tumor data suggests that Neyman's bias may be present because the within exposure stratum association between D and T^* suggests a within exposure stratum association of D and M_d. However, we should restate that an association within strata does not necessary imply that bias is present; it is only when the $D \perp\!\!\!\perp M_d \mid G$ holds that we can conclude that Neyman's bias is not present. Additionally, the study design may contribute to a within stratum association between D and T^* and so the authors suggest that more work is needed to form stronger conclusions regarding the potential presence of Neyman's bias in this study.

As with the result from Test 2, the rejection of the null hypothesis of no Neyman's bias in the stroke-mortality data by Test 3 needs confirmatory analyses. A primary concern is that if the population underlying the measurements in dbSNP, the source of our "true" population MAFs against which we compare the estimate, is significantly different than that composing the study subjects, the type 1 error could be inflated. Since, for many of the SNPs in the data set, the MAF among cases and the MAF among controls did not contain the population MAF, which should be the case as the sample size gets large, there is some evidence of different underlying populations. Another assumption that may

not be satisfied is $P(D < M_d) = 1$. While $P(D < M_d) = 1$ is unlikely to ever be fully satisfied, ischemic stroke is an event with numerous comorbidities and so violations of the assumption may be too large for a valid test (Ostwald et al., 2006; Bots et al., 1997). Lastly, the description of Test 3 showed that the power for detection of bias goes to α as the population prevalence of disease gets small. The implication of this result is that any bias detected when the population prevalence of disease ranges over a relatively small 0.5%-5% is more likely due to unsatisfied assumptions than genuine Neyman's bias. The generous power calculation of 0.06 confirms this belief–it is unlikely that a large proportion of SNPs would have significant p-values, as we have, when there is littler power to detect the bias. It is more likely that the reference population minor allele frequencies are unreflective of the population in the MGH study, which is an assumption that must be satisfied for a valid test.

We did not use Test 1 on the brain tumor and stroke data sets because of an insufficient sample size and insufficient covariates, respectively. The sample size was insufficient in the brain tumor data set because the comparability criterion for Test 1 is more stringent than that for Test 2, so there are only a limited number of pairs of observations that can contribute to estimation of the necessary parameters, especially when overlap between the multivariate random variables $(D \quad T^* \quad M_d)^T$ is minimal. Thus, while Test 2 might be thought of as somewhat removed from testing $D \perp\!\!\!\perp M_d \mid G$ because it tests $D \perp\!\!\!\perp T^* \mid G$ as a proxy, one advantage of Test 2 over Test 1 is that there are fewer restrictions imposed by the comparability criterion, allowing for more flexible use of the data.

Acknowledgements and address.

The authors wish to thank Dr. Deborah Blacker for many helpful comments used in the preparation of this manuscript as well as Drs. Gregory Cairncross and David Louis for use of the brain tumor data. Dr. Guido Falcone provided invaluable support in the preparation of the ischemic stroke dataset. The MGH ischemic stroke dataset was supported by the American Heart Association/Bugher Foundation Centers for Stroke Prevention Research, the National Institute of Neurological Disorders and Stroke, the Deane Institute for Integrative Study of Atrial Fibrillation and Stroke, and the Keane Stroke Genetics Research Fund. Dr. Anderson is supported by a Clinical Research Training Fellowship from the American Brain Foundation. Prof. Rebecca Betensky is supported by the National Institutes of Health grant CA075971. Dr. Swanson was supported by the National Institutes of Health Training Grant T32 NS048005 while the work was completed at Harvard School of Public Health. His current address is GNS Healthcare, 196 Broadway, Cambridge, MA 02139. His email address is dms866@mail.harvard.edu.

Appendix: direction of bias and examples.

We provide a theorem regarding the direction of Neyman's bias under certain modeling assumptions and examples of when Neyman's bias does or does not occur.

Theorem 1 *If G is associated with D such that $OR(t^*) \neq 1$, the distribution of $D \mid (G = 0)$ and $D \mid (G = 1)$ belong to the same location family, $P(X > 0) = 1$, $P(X < t^{**}) > 0$ (where t^{**} is defined as the time between t^* and the first possible presence of disease among the exposed or unexposed), and $X \perp\!\!\!\perp (D \;\; G)^T$, then $OR_{ob}(t^*) \neq OR_{tr}(t^*)$. Specifically, if $D \mid (G = 0)$ is stochastically greater than $D \mid (G = 1)$ (alternatively, stochastically less than) so that exposure is a risk factor for disease (alternatively, protective against disease), then $OR_{ob}(t^*) < OR_{tr}(t^*)$ (alternatively, $OR_{ob}(t^*) > OR_{tr}(t^*)$).*

Proof. Define $\partial F_{D|G=0}(x)/\partial x = f_0(x)$ and $\partial F_{D|G=1}(x)/\partial x = f_1(x)$, and suppose that $f_1(x) = f_0(x-k)$ for some k positive, without loss of generality. Such a scenario corresponds to exposure being protective against disease, though below we will also consider it a risk factor. $f_1(x)$ and $f_0(x)$ are in the same location family. Define $F(x)$ as the cumulative distribution function of X evaluated at x and remember $F(0) = 0$ and $F(t^*) > 0$. Consider the two quantities:

$$\frac{\int_0^{t^*}(1 - F(t^* - x))f_0(x)\partial x}{\int_0^{t^*} f_0(x)\partial x} \quad \text{and} \quad \frac{\int_0^{t^*}(1 - F(t^* - x))f_1(x)\partial x}{\int_0^{t^*} f_1(x)\partial x},$$

which we call the "percent erosion" of $\int_0^{t^*} f_0(x)\partial x$ and $\int_0^{t^*} f_1(x)\partial x$, respectively. Then

$$\begin{aligned}\frac{\int_0^{t^*}(1 - F(t^* - x))f_1(x)\partial x}{\int_0^{t^*} f_1(x)\partial x} &= \frac{\int_0^{t^*}(1 - F(t^* - x))f_0(x-k)\partial x}{\int_0^{t^*} f_0(x-k)\partial x} \\ &= \frac{\int_{-k}^{(t^*-k)}\left[1 - F(t^* - (x+k))\right]f_0(x)\partial x}{\int_{-k}^{(t^*-k)} f_0(x)\partial x}.\end{aligned}$$

Since $F(\cdot)$ a cumulative distribution function and therefore increasing, we have

$$\begin{aligned}\frac{\int_0^{t^*}(1 - F(t^* - x))f_1(x)\partial x}{\int_0^{t^*} f_1(x)\partial x} &= \frac{\int_{-k}^{(t^*-k)}\left[1 - F(t^* - (x+k))\right]f_0(x)\partial x}{\int_{-k}^{(t^*-k)} f_0(x)\partial x} \\ &> \frac{\int_0^{t^*}(1 - F(t^* - x))f_0(x)\partial x}{\int_0^{t^*} f_0(x)\partial x}, \quad (4)\end{aligned}$$

because at every "successive" ∂x in each integral, $1 - F(t^* - (x+k)) \geq 1 - F(t^* - x)$ and there is some $0 < x < t^*$ for which $1 - F(t^* - (x+k)) > 1 - F(t^* - x)$. Thus, the "percent erosion" of $f_0(x)$ will always be greater than that of $f_1(x) = f_0(x-k)$, which is intuitive since $f_1(\cdot)$ is located to the right of $f_0(\cdot)$ and thus subject to the corrosive

effects of $F(\cdot)$ for less "time." Then using the inequality in (4),

$$\begin{aligned} 1 > \left[\frac{\int_0^{t^*}(1-F(t^*-x))f_0(x)\partial x}{\int_0^{t^*} f_0(x)\partial x}\right] \Big/ \left[\frac{\int_0^{t^*}(1-F(t^*-x))f_1(x)\partial x}{\int_0^{t^*} f_1(x)\partial x}\right] \\ = \frac{\int_0^{t^*} f_1(x)\ \partial x\ p}{\int_0^{t^*} f_0(x)\ \partial x\ (1-p)} \times \frac{\int_0^{t^*}(1-F(t^*-x))f_0(x)\ \partial x\ (1-p)}{\int_0^{t^*}(1-F(t^*-x))f_1(x)\ \partial x\ p} \\ = \frac{P(\text{Case, Exposed})}{P(\text{Case, Unexposed})} \times \frac{P(\text{Case, Unexposed, Observed})}{P(\text{Case, Exposed, Observed})}, \end{aligned}$$

which implies that

$$\frac{P(\text{Case, Exposed, Observed})}{P(\text{Case, Unexposed, Observed})} > \frac{P(\text{Case, Exposed})}{P(\text{Case, Unexposed})} \quad \text{and} \quad OR_{ob}(t^*) > OR_{tr}(t^*)$$

since $P(X > 0)$ implies $P(\text{Control, Exposed, Observed}) = P(\text{Control, Exposed})$ and $P(\text{Control, Unexposed, Observed}) = P(\text{Control, Unexposed})$. Again, these inequalities only hold when exposure is protective against disease. When exposure is a risk factor for disease and therefore shifts the mean age of disease onset to the left under the above assumptions,

$$\frac{P(\text{Case, Exposed, Observed})}{P(\text{Case, Unexposed, Observed})} < \frac{P(\text{Case, Exposed})}{P(\text{Case, Unexposed})} \quad \text{and} \quad OR_{ob}(t^*) < OR_{tr}(t^*)$$

using analogous results. So we see that the bias is not toward the null, but in a definite direction depending on model assumptions.

Example 1. Consider $D \mid (G = 1)$ uniform on $(0, 2)$, $D \mid (G = 0)$ uniform on $(0, 1)$, and X uniform on $(0, 3)$, independent of G. Clearly the distributions of disease for exposed and unexposed are not in the same location family in this case, and the model for X corresponds to disease-induced mortality necessarily occurring within 3 times units after disease, D. We need only consider cases when investigating the odds ratio since we assume $P(X > 0) = 1$, implying $P(D < M_d) = 1$. Taking $t^* = 1$,

$$\begin{aligned} \frac{P(\text{Case, Exposed, Observed})}{P(\text{Case, Unexposed, Observed})} &= \frac{\int_0^1 (2/3 + x/3)\ (1/2)\ p\ \partial x}{\int_0^1 (2/3 + x/3)\ 1\ (1-p)\ \partial x} \\ &= \frac{1/2\ \int_0^1 (2/3 + x/3)\ p\ \partial x}{1\ \int_0^1 (2/3 + x/3)\ (1-p)\ \partial x} = \frac{1\,p}{2\,(1-p)} = \frac{P(\text{Case, Exposed})}{P(\text{Case, Unexposed})}. \end{aligned}$$

So we have X independent of exposure status and time of disease-onset, as was the case above, but here $OR_{ob} = OR_{tr}$.

Example 2. Consider again $D \mid (G = 1)$ uniform on $(0, 2)$, and $D \mid (G = 0)$ uniform on $(0, 1)$. However, consider $X \mid (G = 1)$ uniform on $(0, 3)$ and $X \mid (G = 0)$ with density $f_{X|G=0}(x) = 2/3\,(1-x)^2$ on $[0, 1 + (9/2)^{1/3}]$. Again, we need only consider cases when investigating potential bias of the odds ratio since we assume $P(D < M_d) = 1$ so

that controls are not subject to the bias-inducing mortality event. Taking $t^* = 1$,

$$\frac{P(\text{Case, Exposed, Observed})}{P(\text{Case, Unexposed, Observed})} = \frac{\int_0^1 (2/3 + x/3)\,(1/2)\,p\,\partial x}{\int_0^1 (7/9 + 2x^3/9)\,1\,(1-p)\,\partial x}$$
$$= \frac{1/2 \cdot \int_0^1 (2/3 + x/3)\,p\,\partial x}{1\,\int_0^1 (7/9 + 2x^3/9)\,(1-p)\,\partial x} = \frac{1/2\,(5/6)\,p}{1\,(5/6)\,(1-p)} = \frac{1\,p}{2\,(1-p)} = \frac{P(\text{Case, Exposed})}{P(\text{Case, Unexposed})},$$

and so here we have no bias again.

Example 3. Assume the same models of D conditional on G, and suppose $X \mid (G = 1)$ is uniform on $(0, 3)$ and $X \mid (G = 0)$ has density $f_{X|G=0}(x) = 5/2\,(1-x)^4$ on $[0, 1 + 2^{1/5}]$. For the reasons given above, we again only consider cases for investigating the bias of the odds ratio. Taking $t^* = 1$,

$$\frac{P(\text{Case, Exposed, Observed})}{P(\text{Case, Unexposed, Observed})} = \frac{\int_0^1 (2/3 + x/3)\,(1/2)\,p\,\partial x}{\int_0^1 (1/2 + x^5/2)\,1\,(1-p)\,\partial x}$$
$$= \frac{1/2\,\int_0^1 (2/3 + x/3)\,p\,\partial x}{1\,\int_0^1 (1/2 + x^5/2)\,(1-p)\,\partial x} = \frac{1/2\,(5/6)\,p}{1\,(7/12)\,(1-p)} \neq \frac{1\,p}{2\,(1-p)} = \frac{P(\text{Case, Exposed})}{P(\text{Case, Unexposed})},$$

and so here we have bias.

Example 4. Take $D \mid (G = 1)$ with density $f_{D|G=1}(x) = x^2/4$ on $[0, 12^{1/3}]$, $D \mid (G = 0)$ with density $f_{D|G=0}(x) = x/3\,[0, 6^{1/2}]$. Then let $X \mid (G = 1)$ have density $f_{X|G=1}(x) = (2-x)^2/4$ on $[0, 2+4^{1/3}]$ and $X \mid (G = 0)$ be uniform on $[0, 2]$. As before, we need only consider cases when investigating the odds ratio since we assume $P(D < M_d) = 1$ so that controls are not subject to the bias-inducing mortality event. Taking $t^* = 2$,

$$\frac{P(\text{Case, Exposed, Observed})}{P(\text{Case, Unexposed, Observed})} = \frac{\int_0^2 (1/3 + 1/12\,x^3)\,(x^2/4)\,p\,\partial x}{\int_0^2 (x/2)\,x/3\,(1-p)\,\partial x}$$
$$= \frac{(4/9)\,p}{4/9\,(1-p)} = \frac{p\,\int_0^2 (x^2/4)\,\partial x}{(1-p)\,\int_0^2 x/3\,\partial x} = \frac{p}{1-p} = \frac{P(\text{Case, Exposed})}{P(\text{Case, Unexposed})}.$$

Remember that $P(\text{Case, Exposed})/P(\text{Case, Unexposed}) = p/(1-p)$ implies $OR_{tr}(t^*) = 1$ when $P(D < M_d) = 1$, which is assumed from condition 3.

Example 5. On the other hand, we can obtain a biased odds ratio using the same conditional disease models as in the previous example and having $X \mid (G = 1)$ with density $f_{X|G=1}(x) = (2-x)^2/4$ on $[0, 2 + 4^{1/3}]$ and $X \mid (G = 0)$ uniform on $[0, 2]$. We again assume $P(D < M_d) = 1$ from condition 3. Taking $t^* = 2$,

$$\frac{P(\text{Case, Exposed, Observed})}{P(\text{Case, Unexposed, Observed})} = \frac{\int_0^2 (1/2 + 1/16\,x^3)\,(x^2/4)\,p\,\partial x}{\int_0^2 (x/2)\,x/3\,(1-p)\,\partial x} = \frac{p\,(1/2)}{(1-p)\,4/9}$$
$$\neq \frac{(4/9)\,p}{4/9\,(1-p)} = \frac{p\,\int_0^2 (x^2/4)\,\partial x}{(1-p)\,\int_0^2 x/3\,\partial x} = \frac{p}{1-p} = \frac{P(\text{Case, Exposed})}{P(\text{Case, Unexposed})}.$$

References

Anderson, C., Nalls, M., Biffi, A., Rost, N., Greenberg, S., Singleton, A., Meschia, J., and Rosand, J. (2011). The effect of survival bias on case-control genetic association studies of highly lethal diseases. *Circulation. Cardiovascular genetics*, 4(2):188–196.

Austin, M. D., Simon, D. K., and Betensky, R. A. (2013). Computationally simple estimation and improved efficiency for special cases of double truncation. *Lifetime data analysis*, pages 1–20.

Betensky, R., Louis, D., and Cairncross, J. (2003). Analysis of a molecular genetic neuro-oncology study with partially biased selection. *Biostatistics (Oxford, England)*, 4(2):167–178.

Bots, M., Hoes, A., Koudstaal, P., Hofman, A., and Grobbee, D. (1997). Common carotid intima-media thickness and risk of stroke and myocardial infarction: the rotterdam study. *Circulation*, 96(5):1432–1437.

CDC (2012). Prevalence of stroke–united states, 2006-2010. *MMWR*, 61(20):379–382.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–31.

Feigin, V., Lawes, C., Bennett, D., Barker-Collo, S., and Parag, V. (2009). Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. *Lancet neurology*, 8(4):355–369.

Fluss, R., Mandel, M., Freedman, L. S., Weiss, I. S., Zohar, A. E., Haklai, Z., Gordon, E.-S., and Simchen, E. (2012). Correction of sampling bias in a cross-sectional study of post-surgical complications. *Statistics in Medicine*.

Hernan, M. and Robins, J. (in press). *Causal Inference*. Chapman and Hall/CRC.

Hill, G. (2003). Neyman's bias re-visited. *Journal of Clinical Epidemiology*, 56.

Hoeffding, W. et al. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325.

Ino, Y., Betensky, R., Zlatescu, M., Sasaki, H., Macdonald, D., Stemmer-Rachamimov, A., Ramsay, D., Cairncross, J., and Louis, D. (2001). Molecular subtypes of anaplastic oligodendroglioma: implications for patient management at diagnosis. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 7(4):839–845.

Johnston, S., Mendis, S., and Mathers, C. (2009). Global variation in stroke burden and mortality: estimates from monitoring, surveillance, and modelling. *Lancet neurology*, 8(4):345–354.

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 371–412.

Keiding, N. (2006). Event history analysis and the cross-section. *Statistics in medicine*, 25(14):2343–2364.

Martin, E. and Betensky, R. (2005). Testing quasi-independence of failure and truncation times via conditional kendall's tau. *Journal of the American Statistical Association*, 100.

Neyman, J. (1955). Statistics; servant of all sciences. *Science (New York, N.Y.)*, 122(3166):401–406.

Ostwald, S., Wasserman, J., and Davis, S. (2006). Medications, comorbidities, and medical complications in stroke survivors: the cares study. *Rehabilitation nursing : the official journal of the Association of Rehabilitation Nurses*, 31(1):10–14.

Rothman, K., Greenland, S., and Lash, T. (2008). *Modern epidemiology*. Lippincott Williams and Wilkins.

Sackett, D. (1979). Bias in analytic research. *Journal of chronic diseases*, 32(1-2):51–63.

Tsai, W. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika*, 77(1):169–177.

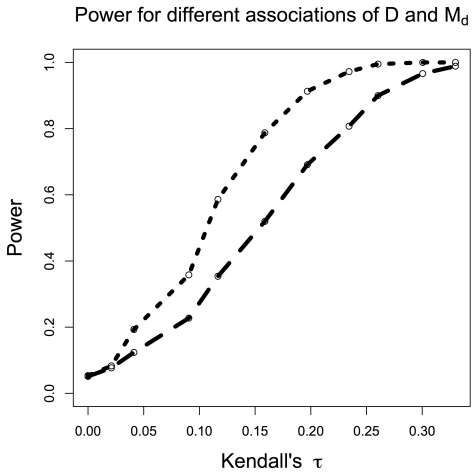

Figure 1: Comparison of power between tests 1 (short dashes) and 2 (long dashes) as a function of the association between D and M_d, measured by Kendall's τ, holding the sample size constant.

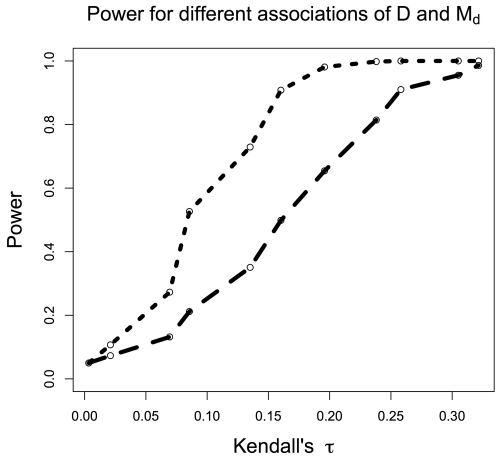

Figure 2: Comparison of power for tests 1 (short dashes) and 2 (long dashes) as a function of the D and M_d as measured by Kendall's τ, holding the number of comparable pairs constant.

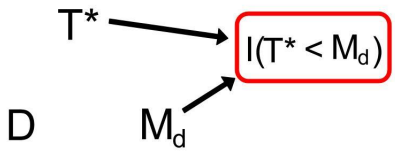

Figure 3: This DAG provides the framework for Test 2. When D is not associated with M_d, there is no association between D and T^*, despite the conditioning event, using rules of DAGs. This figure represents these random variables within each stratum of G.

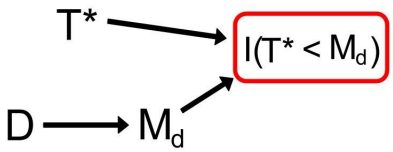

Figure 4: This DAG provides the framework for Test 2. When D is associated with M_d, an association between D and T^* is induced due to the conditioning event using rules of DAGs. This figure represents these random variables within each stratum of G.

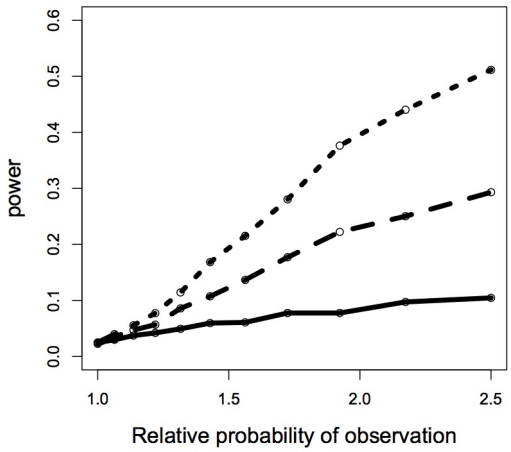

Figure 5: Power for test 3 as a function of $p^*(t^*)$ and the relative probability of observing the unexposed cases versus exposed cases. As the relative probability increases (i.e., it is more likely to observed unexposed cases than exposed cases) as is the case when there are a greater number of mortality-inducing events among the exposed, there is more bias and power. The solid, dashed, and dotted lines represent population prevalences of disease ($p^*(t^*)$) of 0.1, 0.2, and 0.3, respectively.