Published in final edited form as:

Stat Med. 2016 October 30; 35(24): 4398–4412. doi:10.1002/sim.7000.

Who Really Gets Strep Sore Throat? Confounding and Effect Modification of a Time-Varying Exposure on Recurrent Events

Dean Follmann^{*},

Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 5601 Fishers Lane Room 4C11, Rockville, MD 20852

Chiung-Yu Huang, and

Sidney Kimmel Comprehensive Cancer Center and Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205.

Erin E Gabriel

Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 5601 Fishers Lane Room 4C11, Rockville, MD 20852

Abstract

Unmeasured confounding is the fundamental obstacle to drawing causal conclusions about the impact of an intervention from observational data. Typically, covariates are measured to eliminate or ameliorate confounding, but they may be insufficient or unavailable. In the special setting where a transient intervention or exposure varies over time within each individual and confounding is time constant, a different tack is possible. The key idea is to condition on either the overall outcome, or the proportion of time in the intervention. These measures can either eliminate the unmeasured confounding by conditioning or by use of a proxy covariate. We evaluate existing methods and develop new models from which causal conclusions can be drawn from such observational data even if no baseline covariates are measured. Our motivation for this work was to determine the causal effect of *Streptococcus* bacteria in the throat on pharyngitis (sore throat) in Indian schoolchildren. Using our models, we show that existing methods can be badly biased and that sick children who are rarely colonized have a high probability that the *Streptococcus* bacteria is causing their disease.

Keywords

Attributable fraction; Case-crossover design; Pattern mixture models; Probability of causation; Self controlled case series

1. Introduction

The potential for unmeasured confounding makes causal conclusions about an exposure from observational data problematic. A common strategy to attack this problem is to

^{*}Correspondence to: Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 5601 Fishers Lane Room 4C11, Rockville, MD 20852.

measure covariates so that conditional on the covariates, the intervention is effectively assigned at random. In the causal literature there are at least four main methods for controlling confounding in a observational trial to allow for causal inference to be made; confounder adjustment or regression analysis, instrumental variable analysis [1], matching [2], and propensity score adjustment [3]. When potential confounding variables or instrumental variables are available these methods can work well and are straightforward to implement. However, it is not always possible to measure sufficient covariates and so the assumption of no unmeasured confounders is often suspect.

If there is a time-varying intervention, e.g. periodic use of a transiently acting drug, so subjects are observed during both exposed and nonexposed states, a different tack is possible. If intervals of drug exposure are randomly assigned, one can compare outcomes on drug versus off drug within an individual. Conceptually this is no different from a repeated crossover trial. Crossover trials require that the effect of the exposure be achieved by the end of the exposure period and that there is no lingering effect of the exposure beyond the period of exposure. Analysis of intermittently administered interventions for survival type data has been discussed in [4] and [5], while [6] considered the setting of recurrent event data. In observational studies, the time-varying intervention or exposure is not assigned at random and may be related to an individual's underlying risk of an event. Maclure [7] introduced the case-crossover design for such data where the effect of transient exposures on an event are estimated by comparing exposure status just prior to the event to a matched control period in the past; see also [8] and [9]. Navidi [10] extended the case-crossover design approach to recurrent events, while [11] discussed different ways to generalize this method to recurrent events where multiple event periods are matched with multiple control periods. A related approach is the self-controlled case series (SCCS), introduced in [12]. Here each individual follows their own nonhomogeneous Poisson process with some parameters common to all subjects. A parametric multinomial model for the common parameters was derived by conditioning on the number of events for each person. This approach has been used extensively in the study of vaccine safety [13]. Finally, the method of Huang et al. [14] can also be applied to this setting. They specified a semiparametric intensity with a person-specific baseline intensity and some parameters common to all subjects. By applying the pairwise conditional likelihood approach [15] to all pairs of subjects with at least one event, the baseline intensity function is eliminated.

This research was motivated by a study of school children in Vellore, India where throat colonization by Group A *Streptococcus* (GAS) bacteria was measured over time along with events of pharyngitis (sore throat) [16]. Of major interest was to estimate the probability of causation (PC) or the probability that GAS colonization is responsible for pharyngitis in a child who presents with both. The probability of causation was introduced by Levin [17] and discussed by Greenland [18] who cautions against its uncritical use. Because a child's (unmeasured) immune system impacts both colonization by GAS and clinical disease, it is difficult to pretend that GAS colonization is effectively randomized and methods that address confounding are needed.

The idea of this paper is to use the individual follow-up data on the colonization and pharyngitis event rates to act as proxies for the unmeasured confounding. We review

methods that condition on the pharyngitis events and show that the SCCS and pairwise conditioning approach of Huang et al. [14] can be applied to this setting. We introduce a new approach where the colonization rate is used as a proxy for a time-constant unmeasured confounder. The new approach naturally handles effect modification where the probability of causation depends on the unmeasured confounder. The SCCS and Huang et al. methods cannot directly handle such modification. This new approach is similar in spirit to pattern mixture models for missing data where models for the observed (nonmissing) data are created conditional on the patterns of missingness [19]. Simulations show that the SCCS and Huang et al. methods work well if there is no effect modification but can be badly biased otherwise. Using the colonization rate as a covariate proxy for the confounder can provide accurate inference under both settings. For the Vellore schoolchildren study, we show strong effect modification where children who are rarely colonized have a high probability of causation while for frequently colonized children the probability of causation is nearly zero. These results suggest it may be beneficial to tailor antibiotic therapy based on a child's history of colonization.

2. CAUSAL FRAMEWORK

In this section we develop a causal framework that applies to a time-varying binary exposure on a recurrent event under the strong assumption that transient periods of exposure are assigned at random.

To talk of causation, it is natural to think of counterfactual outcomes and so we imagine whether a child would have had the event (pharyngitis) if they were positive for exposure (GAS colonized) and if they were not. For now, we consider the probability of causation for a small interval $[t, t + \Delta]$. Let $Y(Z)$ be the event indicator for a child who has exposure state $Z = 0, 1$ in this interval. Thus of the pair $Y(0), Y(1)$, one is observed, the other is counterfactual, and the observed data is $Y = Y(1)Z + Y(0)(1 - Z)$. There are four possibilities for these pairs of outcomes which are provided in Table 1 along with a meaningful label.

At a given point in time, some children will have $Z = 1$ and others will have $Z = 0$. In this section, we make the strong assumption that the true proportions of Table 1 for the four categories are the same for these two types of children. If the Z values were randomly assigned, this would be automatically satisfied.

Under the assumption that exposure cannot prevent the event or $\theta_{10} = 0$ we write the probability of causation as

$$PC = pr(Z \text{ caused the event} | \text{have the event and } Z=1) = \frac{\theta_{01}}{\theta_{01} + \theta_{11}} \quad (1)$$

$$= \frac{pr(Y=1|Z=1) - pr(Y=1|Z=0)}{pr(Y=1|Z=1)}. \quad (2)$$

The PC measures a causal effect at an individual level and is estimable under the assumption that $\theta_{10} = 0$ since $\text{pr}(Y = 1 | Z = z)$ can be estimated using sample proportions. Note that if, peculiarly, the exposure $Z = 1$ prevented disease in a proportion θ_{10} , then the PC estimated by the right hand side (2) would estimate $(\theta_{01} - \theta_{10})/(\theta_{01} + \theta_{11})$, an underestimate of the true probability of causation.

For our setting, PC is of interest as it additionally provides the probability that effective antibiotics will cure a child of disease, for a child who presents with disease and GAS colonization of the throat. The probability of causation was introduced by Levin [17] and discussed by Greenland [18] for the setting of a long term exposure on a terminal event.

In principle, we could make many tables like Table 1 for the partition of Δ -sized intervals of the time axis and aggregate the estimated θ s and thus PCs. A better way is to postulate a model for the instantaneous risk of the event as a function of current exposure. To proceed, let $Z(t)$ be the indicator of exposure at time t , $N(t)$ the number of events occurring at or before time t . We first assume that, conditioning on $Z(t)$, the recurrent event process $N(t)$ is a nonhomogeneous Poisson process with intensity function

$$\lambda(t) = \lambda_0(t) \exp\{Z(t)\beta_1\}.$$

This specification is known as the proportional intensity model and is the recurrent event analogue to the proportional hazards model for survival data [20, ?]. It follows that in a small interval $[t, t + \Delta)$, the probability of having a pharyngitis event is

$\int_t^{t+\Delta} \lambda(u) du \approx \lambda(t) \Delta$. Thus under the proportional intensity model with random exposure, the probability of causation for a small interval is

$$\begin{aligned} PC &= \frac{\text{pr}\{N(t+\Delta) - N(t) > 0 | Z(t) = 1\} - \text{pr}\{N(t+\Delta) - N(t) > 0 | Z(t) = 0\}}{\text{pr}\{N(t+\Delta) - N(t) > 0 | Z(t) = 1\}} \\ &\approx \frac{\lambda_0(t)e^{\beta_1\Delta} - \lambda_0(t)\Delta}{\lambda_0(t)e^{\beta_1\Delta}} = 1 - \frac{1}{e^{\beta_1}}. \end{aligned}$$

Another important metric is the attributable fraction (AF) or fraction of cases that are attributable to the exposure. Arguing as before, we consider a short period of time and consider the expected number of cases in the current environment (i.e. where $Z = 1$ with probability p over the interval and over the population) compared to an environment where the exposure has been eliminated (e.g., a perfect vaccine which eliminates GAS colonization and therefore disease). We obtain an attributable fraction as

$$AF(p) = \frac{E_{ZC} [\Delta \lambda_0(t) e^{\beta_1 Z(t)}] - \Delta \lambda_0(t)}{E_{ZC} [\Delta \lambda_0(t) e^{\beta_1 Z(t)}]} = 1 - \frac{1}{E_{ZC} [e^{\beta_1 Z(t)}]} = 1 - \frac{1}{e^{\beta_1 p} + (1 - p)}$$

where E_{ZC} denotes the expectation of $Z(t)$ in the current environment. The attributable fraction measures a causal effect on a population. If $\theta_{01} > 0$, the estimate of $AF(p)$ can be less than 0 if, for example, GAS prevented more cases of pharyngitis than it caused. The

reasoning here is analogous to that of a randomized trial where we can imagine the causal population impact of a harmful intervention would be a net increase in disease.

3. UNMEASURED CONFOUNDING

While the methods of the previous section are appropriate when $Z(t)$ is effectively assigned at random, such pseudo-randomization is unlikely to be the case. In our example, children who tend to get colonized may have a different intrinsic risk of pharyngitis than children who tend to be free of colonization. More specifically, the components of the immune system that thwart colonization may be related to the components that fight clinical disease. A simple schematic of how the unmeasured immune system U might impact observable data is given in Figure 1. If the arrow from U to $Z(t)$ were missing, then U can be viewed as a subject specific frailty variable and confounding would not be a concern. If the arrow from U to $N(t)$ were missing, then standard methods that condition on $Z(t)$ could be used to estimate the models. Importantly, this schematic assumes that U is a time-constant quality of a child that impacts both $Z(t)$ and $N(t)$. Relaxation of the assumption of time constancy will be discussed later.

A common approach to deal with confounding is to measure baseline covariates $\mathbf{X}$ that are related to both the intervention/state $Z(t)$ and the outcome $N(t)$. With perfect covariate selection, $\mathbf{X}$ essentially allows identification of U . More specifically we could formalize a model for the intensity function conditional on the history to time t ,

$$\lambda(t) = \lambda_0(t) \exp \left\{ Z(t) \beta_1 + \mathbf{X}' \beta_2 + \mathbf{S}(t)' \beta_3 \right\}, \quad (3)$$

where $\mathbf{X}$ is a vector of baseline covariates that are correlated with U so that $U \approx \mathbf{X}' \beta_2$ and $\mathbf{S}(t)$ is a vector of time-varying covariates that can include functions of the recurrent event history of the child prior to t .

In (3) it is assumed that the effect of changes in $Z(t)$ are immediate and complete. More sophisticated models could allow for gradual effects. For example, a ‘carryover’ impact of a change in $Z(t)$ from 1 to 0 at time t_s could be modeled by replacing $Z(t) \beta_1$ in (3) by, $\{1 - t/(\omega + t_s)\} I\{t \in (t_s, t_s + \omega)\} \beta_1$, so that the impact of the switch gradually diminishes post shift for a period of time of length ω . For simplicity, we will focus on the simpler (3) in the sequel.

In the Vellore pharyngitis study no baseline covariates were measured so we need to develop a different way to address unmeasured confounding. For simplicity, our subsequent development ignores $\mathbf{X}$ and $\mathbf{S}(t)$ though they could be easily added as could gradual effects for switches in $Z(t)$. Our thinking is that with long follow-up we can reliably estimate each child's long term probability of being colonized, $\bar{Z} = \int_0^\tau Z(t) dt / \tau$, and also their rate of pharyngitis, $\bar{Y} = N(\tau) / \tau$, where τ is the duration of follow-up. These can essentially act as separate proxies for U so the unmeasured confounder is increasingly revealed with the passage of time. In the next subsections we formalize these ideas.

3.1. Conditioning on outcome

Let $N(\cdot) = \{N(t); 0 \leq t \leq \tau\}$ and $Z(\cdot) = \{Z(t); 0 \leq t \leq \tau\}$. For notational simplicity, we assume a common follow-up time τ for all subjects. This can be relaxed so subject i has followup time τ_i . In general a model for our data can be informally represented as $f\{N(\cdot), Z(\cdot), U\}$ which can be decomposed as

$$f\{N(\cdot), Z(\cdot), U\} = f\{N(\cdot) | Z(\cdot), U\} \times f\{Z(\cdot), U\}. \quad (4)$$

An exponential model consistent with the strong pseudo-randomization assumption of the previous section can be written in this form by assuming that $f\{N(\cdot) | Z(\cdot), U\} = f\{N(\cdot) | Z(\cdot) | Z(\cdot)\}$, for example by the specification

$$\lambda(t) = \exp\{\beta_0 + Z(t)\beta_1\}. \quad (5)$$

Note that $f\{Z(\cdot), U\}$, i.e. the way the pseudo-randomization of $Z(t)$ depends on U , is arbitrary, can vary idiosyncratically across individuals, and is ignorable under the pseudo-randomization assumption.

A more realistic model that formalizes confounding is to assume that $f\{N(\cdot) | Z(\cdot), U\}$ is given by the proportional intensity model

$$\lambda(t) = \lambda_0(t) \exp\{Z(t)\beta_1 + U\beta_2\}, \quad (6)$$

where U follows some distribution. This model basically allows each child their own proportional deviation $\exp(U\beta_2)$ from the overall rate of pharyngitis $\lambda_0(t) \exp\{Z(t)\beta_1\}$, as if each child is conducting their own randomized trial with periods of exposure ($Z(t) = 1$ or colonization) and control ($Z(t) = 0$, lack of colonization).

Given U and $\{Z(t), 0 \leq t \leq \tau\}$, the number of pharyngitis events at time τ , i.e. $N(\tau)$, is a Poisson random variable with conditional mean

$$E\{N(\tau) | U, Z(t), 0 \leq t \leq \tau\} = \int_0^\tau \lambda_0(s) \exp\{Z(s)\beta_1 + U\beta_2\} ds.$$

Suppose the time interval $[0, \tau]$ is partitioned into K subintervals $\{I_1, \dots, I_K\}$ and assume that the baseline intensity function $\lambda_0(t)$ is piecewise constant within each interval with values $\lambda_{01}, \dots, \lambda_{0K}$. We further assume that the covariate process is constant within each subinterval, that is, $Z(t)$ takes values of $z_1, \dots, z_K$ in the K subintervals. Then, given $(z_1, \dots, z_K)$, $N(\tau)$ is a Poisson random variable with mean

$$\sum_{k=1}^K \lambda_{0k} L_k e^{z_k \beta_1 + U \beta_2},$$

where L_k is the length of the k th interval. Moreover, the number of pharyngitis events with GAS colonization is a Poisson random variable with mean

$$\sum_{k=1}^K \lambda_{0k} L_k e^{z_k \beta_1 + U \beta_2} I(z_k = 1).$$

Let Y_k denote the number of pharyngitis episodes in the k th interval for a single individual. Then the conditional likelihood contribution for that individual's $(Y_1, \dots, Y_K)$ given $N(\tau)$ is proportional to

$$\prod_{k=1}^K \left(\frac{\lambda_{0k} L_k e^{z_k \beta_1 + U \beta_2}}{\sum_{j=1}^K \lambda_{0j} L_j e^{z_j \beta_1 + U \beta_2}} \right)^{Y_k} = \prod_{k=1}^K \left(\frac{e^{\beta_{0k} + z_k \beta_1}}{\sum_{j=1}^K e^{\beta_{0j} + z_j \beta_1}} \right)^{Y_k}, \quad (7)$$

where $\beta_{0k} = \log\{\lambda_{0k} L_k\}$. Equation (7) elegantly eliminates $U\beta_2$ and is a special case of the SCCS method [12, 13]. Asymptotic normality follows from the theory of a parametric multinomial likelihood and sandwich estimates of standard errors can be used to protect against model mis-specification (White 1982). Subjects with $\bar{Z} = 0$ or 1 and subjects who have no events do not provide information to this likelihood.

When $\lambda_0(t)$ is a constant function, that is, $\lambda_{01} = \dots = \lambda_{0K}$, the above model simplifies to a binomial model. The outcome is N_1 the number of events when $Z(t) = 1$. The binomial parameters are $N(\tau)$ the total number of events and the binomial success probability is given by

$$\frac{\bar{Z} / \left((1 - \bar{Z}) e^{\beta_1} \right)}{\bar{Z} / \left((1 - \bar{Z}) e^{\beta_1} + 1 \right)}, \quad (8)$$

where $\bar{Z} = \sum_{k=1}^K L_k I(z_k = 1) / \sum_{j=1}^K L_j$ is the proportion of time with $Z(t) = 1$.

In these conditional approaches, we assumed a piecewise constant baseline intensity $\lambda_0(t)$. We can completely eliminate an arbitrary nuisance intensity function with a clever kind of conditioning applied to (6). Following [14], we consider a pair of event times from person i and k , say t_{ij} and $t_{k\ell}$. Define a pair of event times as comparable if both persons i and k are under observation at times t_{ij} and $t_{k\ell}$ that is $\max(t_{ij}, t_{k\ell}) \leq \min(\tau_i, \tau_k)$, where τ_i, τ_k are the follow-up duration for persons i and k , respectively. Define the indicator of this event as $I_{ik} = \max(t_{ij}, t_{k\ell}) \leq \min(\tau_i, \tau_k)$. By conditioning on the order statistics of comparable event times, the pairwise pseudo-likelihood contribution for that pair reduces to

$$\frac{1}{1 + \exp(\{Z_i(t_k) + Z_k(t_i)\} - \{Z_i(t_i) + Z_k(t_k)\} \beta_1)}.$$

Note that both $\lambda_0(t)$ and $U\beta_2$ are eliminated from (6) by this pairwise conditioning. Estimation of β follows from maximizing the pairwise pseudo-likelihood

$$\prod_{i < k} \prod_{j=1}^{m_i} \prod_{\ell=1}^{m_k} \left\{ \frac{1}{1 + \exp([\{Z_i(t_{k\ell}) + Z_k(t_{ij})\} - \{Z_i(t_{ij}) + Z_k(t_{k\ell})\}] \beta_1)} \right\}^{I_{ik}}, \quad (9)$$

where m_i is the number of events for person i and t_{ij} is the j th event time for person i . Huang et al. [14] proved asymptotic normality and provided sandwich estimates of variance to protect against model mis-specification [14].

A more general model would allow for effect modification where the impact of $Z(t)$ on the risk of pharyngitis might vary with U . This might make sense for our example if children who tended to be colonized had a different probability of causation from children who tended to not be colonized. A model reflecting this interaction between U and the intervention is

$$\lambda(t) = \lambda_0(t) \exp\{Z(t)\beta_1 + U\beta_2 + Z(t)U\beta_3\}. \quad (10)$$

Unfortunately, neither conditioning on $N(\tau)$ nor pairwise conditioning eliminates U from the resultant model. A simple method to address such effect modification will be introduced in the next section.

3.2. Conditioning on $\bar{Z}$

Another kind of decomposition leads to a different class of models. We can informally write

$$\begin{aligned} f\{N(\cdot), Z(\cdot), U\} &= f\{N(\cdot) | Z(\cdot), U\} \times f\{Z(\cdot), U\} \\ &= f\{N(\cdot) | Z(\cdot), \bar{Z}, U\} \times f\{Z(\cdot), \bar{Z}, U\}. \end{aligned}$$

We postulate that $f\{N(\cdot) | Z(\cdot), \bar{Z}, U\} = f\{N(\cdot), \bar{Z}\}$. This can be justified by either thinking that U is unnecessary once $\bar{Z}$ is known, or as an approximation if $\bar{Z}$ is a good proxy for U . Detailed justification of the latter is given in the appendix. In this model $\bar{Z}$ is the confounder while U is an unobserved reflection of it. Of course $\bar{Z}$ is determined by $Z(\cdot)$ so the last assumption is equivalent to assuming $f\{N(\cdot) | Z(\cdot), U\} = f\{N(\cdot) | Z(\cdot)\}$. An example of this decomposition is to assume that $f\{N(\cdot) | Z(\cdot) | \bar{Z}\}$ is given by

$$\lambda(t) = \lambda_0(t) \exp\{Z(t)\beta_1 + h(\bar{Z})\beta_2 + Z(t)h(\bar{Z})\beta_3\}. \quad (11)$$

where $h(\cdot)$ is a known function such as the identity or logit. This allows the impact of $Z(t) = 1$ to be different for children with different $\bar{Z}$ s. With $\beta_3 = 0$, (11) is similar to (10) but with $\bar{Z}\beta_2$ taking the place of $U\beta_2$. Note that as in the previous section, this decomposition obviates specification of $f\{Z(\cdot), \bar{Z}, U\}$ —the method by which U determines $Z(\cdot)$. The model can be estimated as described in Lin et al. (2000). As before, sandwich estimators of the variance

can be used to effectively weaken the assumption that conditional on $\bar{Z}(\cdot)$, event times are independent.

In the proportional intensity model defined by (11) the risk of an event at time t depends on $\bar{Z}$ which depends on data measured after time t . This seems odd, but this kind of modeling parallels pattern mixture models for missing data [19]. In such models, outcomes are modeled conditional on patterns of missingness (e.g. early dropout versus late dropout), and then an overall model for outcome is obtained by averaging over the patterns of missingness. Similar to our setting, the patterns of missingness are discerned in the future, relative to the outcomes being modeled. Equation (11) can lead to useful inference. First, we can always marginalize over the distribution of $\bar{Z}$ to obtain unconditional inference. Using this maneuver, the (unconditional) probability of causation is given by

$$PC = \int 1 - \frac{1}{e^{\beta_1 + h(\bar{Z})\beta_3}} dF(\bar{Z}),$$

where $F(\bar{Z})$ can be estimated by the empirical distribution function of $\bar{Z}$. Second, the dependence of PC on $\bar{Z}$ may be of scientific interest. For example, with $\bar{Z}$ the proportion of time colonized with GAS, a $\beta_3 < 0$ might suggest a commensal or benign relationship between the pathogen (GAS) and host (child) for larger $\bar{Z}$ and a parasitic or harmful relationship between pathogen and host for small $\bar{Z}$. Having the relationship between GAS and host change with $\bar{Z}$ would refute the idea that GAS is always parasitic. Finally, once the data are collected, the best estimate of the future PC for a child with a historically measured $\bar{Z}_i$ would be the conditional PC

$$PC(\bar{Z}) = 1 - \frac{1}{e^{\beta_1 + h(\bar{Z}_i)\beta_3}}.$$

In addition to (11) being a model in its own right, it can be viewed as a kind of approximation to (10) for a time-constant confounder U . Under weak conditions, for example as given in the appendix, one can show that as $\tau \rightarrow \infty$, $h(\bar{Z}_i) \rightarrow h(z_i) = U_i$. Thus with a long followup, $h(\bar{Z})$ becomes a proxy for U .

This approximation is predicated on the idea that τ is ‘large enough’. If τ is not large enough, the approximation may be poor and one can explore different corrections. Under the assumption of a time-constant confounder and weak conditions on the distribution of sojourn times, we have $h(\bar{Z}_i) \approx N\{U_i, \omega^2(U_i)\}$ for large τ (see Appendix). Put another way we have

$$h(\bar{Z}_i) \approx U_i + e_i \quad (12)$$

where e_i has a normal distribution with variance $\omega^2(U_i)$. Juxtaposing (12) with (11) reveals a classic Gaussian measurement error model where we observe the error-prone $h(\bar{Z}_i)$ instead of the true covariate U_i . Measurement error corrections could be entertained if we could estimate $\omega^2(U_i)$, but this is not estimable without further assumptions. Another approach

would be to use the decomposition (4) with $f\{Z(\cdot), |Z(\cdot), U\}$ given by (10) and $f\{Z(\cdot) | U\} = f\{Z(\cdot) | U\} f(U)$ where $f(U)$ is Gaussian and $f\{Z(\cdot) | U\}$ follows some parametric process. This development parallels shared parameter model for missing data (see e.g. [21] and [22]). Here the shared parameter is U_j and it governs the observable event times and $\bar{Z}_i$. In shared parameter models for missing data, the shared parameter governs both the outcomes and the missingness indicators. Formal development of these approaches is beyond the scope of this paper but could be interesting future work.

4. SIMULATIONS

In this section we evaluate 6 different methods of estimation under different scenarios by simulation.

1. Naive: Here we fit the simple proportional intensity model with no confounding,

$$\lambda(t) = \lambda_0(t) \exp\{Z(t) \beta_1\}.$$

2. Binomial: We assume the proportional intensity model has a constant baseline intensity function modulated by an unmeasured confounder U ,

$$\lambda(t) = \exp\{\beta_0 + Z(t) \beta_1 + U \beta_2\},$$

and condition on the number of events. This leads to a binomial likelihood with parameter β_1 from equation (8) and is a simple SCCS approach.

3. Multinomial: We assume a piecewise constant exponential baseline rate function modulated by an unmeasured confounder U ,

$$\lambda(t) = \exp\left\{\sum_{k=1}^5 \beta_{0k} I(t \in I_k) + Z(t) \beta_1 + U \beta_2\right\}$$

where I_k defines the intervals $(0, 0.2), (0.2, 0.4), \dots, (0.8, 1)$.

Conditioning on $N(\tau)$ leads to a multinomial likelihood with 4 β_{0k} parameters and β_1 . This method is a more sophisticated SCCS approach.

4. Pair: We maximize the pairwise pseudo-likelihood given by (9) to estimate β_1 . This eliminates $\lambda_0(t)$ and U from the proportional intensity model with arbitrary baseline intensity

$$\lambda(t) = \lambda_0(t) \exp\{Z(t) \beta_1 + U \beta_2\}.$$

5. $\bar{Z}$: We fit a proportional intensity model where the intervention interacts with $\bar{Z}$

$$\lambda(t) = \lambda_0(t) \exp\{Z(t) \beta_1 + \bar{Z} \beta_2 + \bar{Z} Z(t) \beta_3\},$$

This approximates both (6) (with $\beta_3 = 0$) and (10).

6. Omni: We fit the correct “omniscient” proportional intensity model that has knowledge of U

$$\lambda(t) = \lambda_0(t) \exp\{Z(t)\beta_1 + U\beta_2 + UZ(t)\beta_3\}.$$

Based on these regression parameter estimates, we evaluate estimates of PC which equals $1 - 1/\exp(\hat{\beta}_1)$ for the first four methods of estimation.

For Method 5, $\widehat{PC}$ equals $E_Z \left\{ 1 - 1/\exp(\hat{\beta}_1 + Z\hat{\beta}_3) \right\}$ where $E_Z(\cdot)$ is over the empirical distribution of the sample $\bar{Z}_{iS}$. For the omniscient Method 6, $\widehat{PC} = E_U \left\{ 1 - 1/\exp(\hat{\beta}_1 + U\hat{\beta}_3) \right\}$ where $E_U(\cdot)$ is over the empirical distribution of the U s.

To generate data we use the decomposition

$$f\{N(\cdot), Z(\cdot), U\} = f\{N(\cdot) | Z(\cdot), \bar{Z}, U\} \times f\{Z(\cdot), \bar{Z} | U\} \times f(U).$$

We specify

- $f(U)$ is given by $U \stackrel{D}{=} N(0, \sigma_U^2)$
- $f(Z), \bar{Z} | U$ is given by $Z(t) = 1$ if $t \sim \exp(\mu + U + \epsilon) / \{1 + \exp(\mu + U + \epsilon)\}$ and $Z(t) = 0$ otherwise, where $\epsilon \stackrel{D}{=} N(0, \sigma_e^2)$, and thus $\bar{Z} = \int_0^1 Z(s) ds$
- $f(N(\cdot) | Z(\cdot), U)$ is determined by the proportional intensity model

$$\lambda(t) = \lambda_0(t) \exp\{\beta_1 Z(t) + U\beta_2 + UZ(t)\beta_3\}$$

For our first scenario, we assume $\lambda_0(t) = \exp(\beta_0)$. We let $\mu = -2$ and $\mu = 0$, the latter induces a symmetry in the distribution of $\bar{Z}$ around 0.5, while $\mu = -2$ better approximates the Vellore study. We consider three settings for $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$: (0, 1, 0, 0), (0, 1, 1, 0), and (0, 1, 1, -1); these correspond to no confounding, confounding, and confounding with interaction. For each β we vary $\sigma_e, \sigma_U = (2, 1), (2, .5), (1, 1), (1, .5), (.5, 1)$, and $(.5, .5)$. Decreasing σ_U tends to reduce the impact of unmeasured confounding while decreasing σ_e tends to make $\bar{Z}$ a better proxy for U as would happen with a larger study. However, smaller σ_U, σ_e induce less variation in $\bar{Z}$ which makes it harder to estimate the coefficients associated with $\bar{Z}$ in (11), namely β_2, β_3 .

Table 2 presents selected results for $\mu = -2$. For all simulation results we use robust estimators of the mean and variance (median and scaled median absolute deviation) due to occasional outlier estimates of PC , especially for the simpler methods. The robust estimates more clearly reveal performance of the estimators.

For Setting 1, with no confounding, $\beta = (0, 1, 0, 0)$, we are simulating a randomized trial and all methods are unbiased for PC. For $(\sigma_e, \sigma_U) = (1, 1)$ the naive and omniscient methods have the smallest MSE while for $(\sigma_e, \sigma_U) = (0.5, 0.5)$ the $\bar{Z}$ method has the largest MSE. For Setting 2, $\beta = (0, 1, 1, 0)$, U has an impact in terms of confounding but no effect modification. Here the naive method is quite biased for all $(\sigma_U, \sigma_e) = (1, 1)$, while the other methods are all unbiased. The proportional rate $\bar{Z}$ method is unbiased but has larger variance than the binomial, multinomial, and pairwise methods. Note that the binomial, multinomial, and pairwise methods are properly specified here while the proportional intensity model with $\bar{Z}$ is a kind of poor man's approximation for U and is unnecessarily estimating an interaction term. For $(\sigma_U, \sigma_e) = (.5, .5)$ the naive and $\bar{Z}$ methods are mildly biased and the $\bar{Z}$ method has the largest MSE. For setting 3, $\beta = (0, 1, 1, -1)$, U is both a confounder and an effect modifier. Here we see that only the $\bar{Z}$ and omniscient methods are unbiased. Interestingly for $(\sigma_U, \sigma_e) = (1, 1)$ the MSE of the $\bar{Z}$ and omniscient methods are similar.

Over all settings, the pairwise conditional model of Huang et al. looks similar to the SCCS multinomial model in terms of bias and variance. In these simulations with an exponential baseline rate function, methods that allow for nonconstant or arbitrary baseline rate functions are unnecessarily flexible. The simple proportional intensity model with $\bar{Z}$ is unbiased for nearly all scenarios and the only reliable choice for Scenario 3. In summary, the pairwise conditional approach of Huang et al. is preferred unless scenario 3 holds in which case the $\bar{Z}$ model is necessary. Simulations with $\mu = 0$ reveal very similar results for the means of the estimates but generally smaller variances as more events result from $\mu = 0$. Simulations where the $Z(t) = 1$ epochs are randomly assigned over time (instead of always at the start) reveal similar behavior. The behavior for the unreported σ_U, σ_e s is similar.

For our next set of simulations we set $\lambda_0(t)$ to be piecewise constant, $-0.45, -0.15, 0.15, 0.45, 0.75$, over 5 intervals of equal length. We set $\bar{Z}^* = \lceil 5\bar{Z} \rceil / 5 + 1 * B$, where $\bar{Z}$ was generated as above, $\lceil \cdot \rceil$ is the greatest integer function, and B is Bernoulli(1/2). Thus $\bar{Z}^*$ is a kind of discretized version of $\bar{Z}$ that can equal 0, 0.2, . . . , 1, and $\bar{Z}^*$ is a poorer proxy for U than $\bar{Z}$. Based on $\bar{Z}^*$ we either randomly assigned the epochs of 0 and 1 over the 5 intervals, or placed all the 1s over the first intervals. For this latter scenario $Z(t)$ tends to be conflated with the piecewise $\lambda_0(t)$.

Table 3 reports selected results for $\mu = -2$ and conflation of $Z(t) = 1$ with followup time. We see again that the proportional intensity model does best for Scenario 3 while the pairwise and multinomial models do best for Scenario 2. The pairwise and multinomial methods fail quite spectacularly if there is differential confounding i.e. $\beta_3 \neq 0$, but are unbiased when $\beta_3 = 0$. The behavior for the unreported σ_U, σ_e s is similar. When we re-ran the simulation with epochs of $Z(t) = 1$ randomized, the performance of the methods were similar to Table 2. Simulations with $\mu = 0$ reveal generally very similar results for the means of the estimates but generally smaller variances as more events result when $\mu = 0$;

In practice one would pick a method that is informed by inspection of the data, thus one might fit a model which allows for interaction with $\bar{Z}$ or functions of $\bar{Z}$. If an interaction model is not necessary, then the pairwise conditional model would be appealing as it eliminates arbitrary subject specific baseline intensity functions with little loss of efficiency.

5. ANALYSIS OF THE VELLORE STUDY

A World Health Organization report estimates that there are over 616 million new cases per year of GAS pharyngitis, of which over 550 million occur in less developed countries. Pharyngitis is most frequently due to viruses, but several bacteria, including Group A streptococci (GAS), persist as a common cause of pharyngitis even in the era of antibiotics. For a child who presents with GAS and pharyngitis, it is not at all certain that the GAS is responsible as the GAS may be incidental to a viral cause. Because there can be serious complications following untreated streptococcal pharyngitis such as rheumatic heart disease and toxic shock syndrome, the standard practice is to treat with penicillin or other antibiotics, even though there is recognition that some unknown proportion of these will be treated unnecessarily. Overuse of antibiotics is a serious worldwide public health concern as it can lead to the development of resistant strains of bacteria which are not treatable by current antibiotics [23, 24]. Thus it is of substantial interest to estimate the probability of causation and possibly tailor antibiotic therapy to those with a high probability of a bacterial cause.

To address these issues, a study of 307 school children in a rural area near Vellore, India was conducted between March 2002 and March 2004 [16]. The enrolled children were followed longitudinally in two different ways. Cases of pharyngitis were identified weekly and throat swabs were obtained on those with pharyngitis to identify the presence of GAS. We call this the pharyngitis data. Children with GAS positive pharyngitis were treated with a brief course of penicillin. We obtained on the student volunteers to determine the prevalence of GAS. We call this the carriage data. Statistically, GAS colonization is a time-varying covariate with pharyngitis the recurrent event.

Overall, 218 of the 307 children had at least one episode of pharyngitis. The median number of episodes was 2 with a maximum of 12. The median colonization rate was 0.08 with 25th and 75th percentiles of 0 and 0.17, respectively; 125 children were never colonized while 3 were always colonized. Table 4 provides a tally of the data while Figure 1 provides data trajectories for 22 selected children. Notable are gaps during school holidays and the enrollment of additional children to make up for those who left at the end of a school year.

We begin our analysis by fitting the simple proportional intensity model

$$\lambda(t) = \lambda_0(t) \exp\{Z(t)\beta_1\}.$$

The carriage data does not provide continuous measurement of $Z(t)$ so we use the value from each monthly visit to specify $Z(t)$ until the next monthly visit, except at event times when the GAS readout from the pharyngitis visit is used.

Under this naive model we obtain an estimate of the probability of causation as $1 - 1/\exp(0.25) = 0.22$ with a 95% bootstrap confidence interval of (0.04, 0.38). This estimand has a casual interpretation if $Z(t)$ is essentially randomly assigned, however this assumption is suspect. We next condition on the total number of events from each child, forming the binomial model with offset of $\log\{\bar{Z}(1 - \bar{Z})\}$ as given in (8). This requires a constant

baseline intensity function and eliminates confounding due to a main effect. Only children with at least 1 event and both positive and negative carriage visits contribute to the likelihood. This filtering results in reducing the 307 children to 136 and the number of events from 640 to 446. The median $\bar{Z}$ over the 136 selected children is 0.15. The resultant PC is 0.14 with a 95% bootstrap confidence interval of $(-0.08, 0.35)$. We also estimate PC using the pairwise approach which allows for an arbitrary baseline intensity; this gives a PC of 0.12 with a 95% bootstrap confidence interval of $(-0.18, 0.32)$. Of course a probability cannot be less than zero. The negative values here reflect the sampling variability in the estimate of PC. If the estimated PC were reliably negative it would call into question the assumption that GAS colonization cannot prevent disease (i.e. that $\theta_{10} = 0$ of Table 1).

Our final analysis allows for different effects of $Z(t)$ for children who are either above or below the overall median colonization rate. We fit the proportional intensity model

$$\lambda(t) = \lambda_0(t) \exp \left[Z(t) \beta_1 + I \left\{ \bar{Z} > \text{median}(\bar{Z}) \right\} \beta_2 + I \left\{ \bar{Z} < \text{median}(\bar{Z}) \right\} Z(t) \beta_3 \right], \quad (13)$$

where $\text{median}(\bar{Z}) = .08$. This model uses all children including those with no events. While it is an easy-to-interpret model in its own right, it can also be viewed as a direct proxy for the model

$$\lambda(t) = \lambda_0(t) \exp \left[Z(t) \beta_1 + I \left\{ U > \text{median}(U) \right\} \beta_2 + I \left\{ U < \text{median}(U) \right\} Z(t) \beta_3 \right], \quad (14)$$

under the weak assumption that $E(\bar{Z}_i) \rightarrow g(U_i)$ for any monotone $g(\cdot)$. In other words, the covariate choice of U as above or below the median obviates concern about what function maps U_i to $\bar{z}_i$.

For this model, the three β coefficients are, respectively .98, .33 and $-.99$ with $p < .005$ for each based on Wald tests with sandwich-based standard errors. This indicates strong effect modification and casts doubt on the estimates of the previous paragraph. The estimated probabilities of causation are 0.62 with a 95% CI of $(0.40, 0.76)$ and -0.01 with a 95% CI of $(-0.34, 0.22)$, for children who are colonized less than or more than the median, respectively. Thus for children who are colonized rarely, the novel insult of GAS has a high probability of causing the pharyngitis, while for children who have repeated colonizations the familiar GAS colony in the throat is unlikely to cause the pharyngitis. In practice, if one knew a child's history of colonization the PC conditional on $\bar{Z}$ could be used to guide therapy with antibiotics strongly recommended for children with a rare bout of colonization. These estimates can also be used to provide an overall estimate of the probability of causation for a randomly selected child without knowledge of past colonization status. The marginalized PC is $.62 \times 0.5 - 0.01 \times 0.5 = 0.31$ which has 95% bootstrap confidence interval of $(0.12, 0.44)$.

For (13) to be an approximation to (14) requires that τ be large enough. We reasoned that if estimates with $\tau/2$ follow-up were close to those with τ follow, then we would have evidence that τ were large enough. We thus created two datasets, one for each of the two school years,

and reran the analysis. This provides us with two estimates, each based on approximately $\tau/2$ follow-up. We obtained marginalized PC estimates of .33 for year 1 and .27 for year 2, both close to the .31 based on the entire dataset. We also fashioned a test to see if either were statistically far from the .31 based on the entire dataset. We simulated a null distribution by randomly selecting half the children (and their entire followup) and calculating the marginalized PC. Under the null that $E\{\widehat{PC}(\tau/2)\} = E\{\widehat{PC}(\tau)\}$, this simulated distribution should approximate the distribution of an estimate of $\widehat{PC}(\tau/2)$, in this dataset as both are based on half the total information. The year 1 and year 2 marginalized PCs fall at the 38th and 67th percentiles of this distribution, corresponding to two-sided p-values of .76 and .66, respectively. We thus have evidence that τ (or even $\tau/2$) is large enough for the approximation of (13) to (14) to be accurate.

Finally, we performed a sensitivity analysis for the assumption of “no carryover” i.e. that the effect of $Z(t)$ switching from 1 to 0 is immediate and complete. We created an additional time-varying covariate which was 1 for a window of length m months following a 1 to 0 switch and 0 otherwise. For $m=2$ and 4, the effect of this covariate was near zero and there was minimal change in the estimates of the other β s.

Another estimand of interest is the attributable fraction. This estimates the proportion reduction in clinical pharyngitis in a new world where clinical pharyngitis due to GAS is eliminated. In this development we assume that GAS colonization cannot prevent disease. For the proportional intensity model with an interaction term for $Z(t)$ and $\bar{Z}$ the construction is somewhat tricky. The subtlety here is how to calculate the risk of pharyngitis for a child who had $\bar{Z}=p$ in the Vellore study in this new world. Because $\bar{Z}$ is essentially a proxy for an unmeasured confounder U , it is an intrinsic quality of the child and from (13) her risk of pharyngitis at time t when she is uncolonized is

$$\lambda_0(t) \exp[\beta_2 I\{p > \text{median}(p)\}].$$

It follows that the attributable fraction can be expressed as

$$AF = 1 - \frac{E_{\bar{Z}}[\Delta\lambda_0(t) \exp\{\beta_2 I(\bar{Z} > \text{median}(\bar{Z}))\}]}{E_{ZC}[\Delta\lambda_0(t) \exp\{\beta_1 Z(t) + \beta_2 I(\bar{Z} > \text{median}(\bar{Z})) + \beta_3 I(\bar{Z} > \text{median}(\bar{Z})) Z(t)\}]}.$$

Here $E_{\bar{Z}}$ is the expectation over the distribution of $\bar{Z}$, while E_{ZC} is the expectation of $Z(t)$ in the current environment. We can estimate AF as

$$1 - \frac{\sum_{i=1}^n \exp\{\hat{\beta}_2 I(\bar{z}_i > \text{median}(\bar{z}_i))\} / n}{\sum_{i=1}^n [\bar{z}_i \exp\{\hat{\beta}_1 + I(\bar{z}_i > \text{median}(\bar{z}_i))(\hat{\beta}_2 + \hat{\beta}_3)\} + (1 - \bar{z}_i) \exp\{I(\bar{z}_i > \text{median}(\bar{z}_i))\hat{\beta}_2\}] / n}.$$

An attributable fraction is of policy interest as it can estimate the impact on pharyngitis disease of a perfectly effective vaccine with perfect community uptake. We estimate the AF

as $-.01$ with a 95% bootstrap confidence interval of $(-.19, .14)$. As mentioned earlier, the AF can be truly less than 0 if, on average, GAS prevents more cases of pharyngitis than it causes. An estimated AF can be less than 0 due to sampling variability. In the Vellore study, 530 of 640 cases of pharyngitis had no GAS present. Thus over 80% of pharyngitis cases could not be caused by GAS and the fraction of pharyngitis cases attributable to GAS is near zero with uncertainty about the estimate reflected in the confidence interval. This analysis suggests that a perfect vaccine would have little impact on pharyngitis disease in this population.

6. DISCUSSION

This paper has examined how different methods can be used to draw causal conclusions for a transient exposure on recurrent event data in the presence of a time-constant confounder. We considered methods that conditioned on the total number of events, the timing of pairs of events, and the overall intervention rate for each person. Without effect modification, but allowing for confounding and arbitrary baseline intensity the pairwise conditional approach of [14] is appealing as it eliminates the nuisance confounding and the nuisance baseline intensity function. With effect modification, a simple proportional intensity model that allows for an interaction with $\bar{Z}$ —a proxy for the unmeasured confounder—does best, while other methods can be badly biased. The key trick underling the proportional intensity model with interaction is to use the future $Z(t)$, via $\bar{Z}$ as a proxy for unmeasured confounding. This trick is similar to pattern mixture models for missing data.

For simplicity (and necessity for the Vellore study) our development avoided the incorporation of other covariates, however, they can be easily incorporated into the methods. These may serve as additional proxies for unmeasured confounding or be of scientific interest by themselves for the proportional intensity model. Methods that condition on outcome (binomial, multinomial, pairwise) eliminate time-constant covariates, however, time-varying covariates are not eliminated and can be used in these conditional methods, or generalizations of them. Our development assumed that the data were conditionally independent, given the conditioning variables (for example, $Z(t)$ and/or $N(\tau)$). In practice there might be auto-correlation in the outcomes even given these conditioning variables. To address this, time-varying functions of the outcome history might be incorporated as additional covariates in our models to weaken the assumption of conditional independence.

We made the strong assumption that the confounder did not vary with time. In practice, this assumption can be weakened by partitioning the data into sets within which the confounder is approximately time-constant, estimating parameters within these sets and then combining the estimates using inverse variance weights. For example, we could break the data in half by calendar time and stratify. Or for each person we could partition their time axis into 4ths, calculate $\bar{Z}$ for each and then stratify by the 2 high $\bar{Z}$ s versus the 2 low $\bar{Z}$ s. This would allow for different probabilities of causation during transient epochs where $Z(t)$ tended to be relatively common for an individual, or relatively rare for an individual. Further development of such approaches is left to future research.

Acknowledgments

Huang's work was sponsored by National Institutes of Health grant 1R01CA193888.

APPENDIX

In this appendix we sketch out a class of models where $h(\bar{Z}) \rightarrow U$ as $\tau \rightarrow \infty$. Rather than working directly with the process $\{Z(s); 0 \leq s \leq \tau\}$ consider the representation of $\{Z(s); 0 \leq s \leq \tau\}$ as the alternating sojourn times in state 0 and 1, say $T_1^0, \dots, T_{N^0(\tau)}^0$ and $T_1^1, \dots, T_{N^1(\tau)}^1$ respectively, where $N^z(\tau)$ is the random number of sojourn times in state z by time τ . We assume that $E(T_j^z) = \mu_z(U)$, $\text{var}(T_j^z) = \sigma_z^2(U)$, that sojourn times are independent conditional on U .

We have $N^z(\tau) \rightarrow \infty$ as $\tau \rightarrow \infty$. Consider the random sum

$$\bar{T}^z(\tau) = \sum_{j=1}^{N^z(\tau)} \frac{T_j^z}{N^z(\tau)},$$

which is random because $N^z(\tau)$ is random. Now by the law of large numbers as $\tau \rightarrow \infty$, we have

$$\bar{T}^z(\tau) \rightarrow \mu_z(U),$$

in probability as $\tau \rightarrow \infty$.

By the central limit theorem for iid random variables we have

$$\sqrt{N^z(\tau)} \left\{ \bar{T}^z(\tau) - \mu_z(U) \right\} \rightarrow N \left\{ 0, \sigma_z^2(U) \right\}.$$

Next, consider the proportion of time that the generic person is in state 1 over the interval $(0, \tau)$. This is

$$\bar{Z}(\tau) = \frac{\sum_{j=1}^{N^1(\tau)} T_j^1}{\sum_{j=1}^{N^1(\tau)} T_j^1 + \sum_{j=1}^{N^0(\tau)} T_j^0} = \frac{\sum_{j=1}^{N^1(\tau)} T_j^1 / \bar{N}(\tau)}{\sum_{j=1}^{N^1(\tau)} T_j^1 / \bar{N}(\tau) + \sum_{j=1}^{N^0(\tau)} T_j^0 / \bar{N}(\tau)},$$

in distribution as $\tau \rightarrow \infty$ where $\bar{N}(\tau) = \{N^1(\tau) + N^0(\tau)\}/2$. Note that $|N^1(\tau) - N^0(\tau)|$ is either 0 or 1 for all τ as the sojourn times alternate. We can thus show that the asymptotic results for

$$\bar{T}^z(\tau) = \sum_{j=1}^{N^z(\tau)} \frac{T_j^z}{N^z(\tau)}$$

are the same as the asymptotic results for

$$\sum_{j=1}^{\overline{N(\tau)}} \frac{T_j^z}{\overline{N(\tau)}}.$$

Thus as $\tau \rightarrow \infty$,

$$\overline{Z} \rightarrow \frac{\mu_1(U)}{\mu_1(U) + \mu_0(U)} = \overline{z}(U),$$

in probability.

By the delta method we obtain

$$\sqrt{\overline{N(\tau)}} \{ \overline{Z} - \overline{z}(U) \} \rightarrow N \left(0, \tau^2(U) = \frac{\mu_0(U)^2}{\{\mu_0(U) + \mu_1(U)\}^4} \sigma_1^2(U) + \frac{\mu_1(U)^2}{\{\mu_0(U) + \mu_1(U)\}^4} \sigma_0^2(U) \right),$$

in distribution as $\tau \rightarrow \infty$.

Define $h()$ as satisfying

$$U = h \left(\frac{\mu_1(U)}{\mu_1(U) + \mu_0(U)} \right).$$

By another application of the delta-method we have

$$\sqrt{\overline{N(\tau)}} \{ h(\overline{Z}) - U \} \rightarrow N \left\{ 0, \{ h'(U) \}^2 \tau^2(U) \right\} \quad (15)$$

in distribution as $\tau \rightarrow \infty$.

Based on (15) we obtain the classical measurement error model which can be written as

$$h(\overline{Z}) \approx U + \epsilon$$

where ϵ is distributed $N[0, \omega^2(U) = \{ h'(U) \}^2 \tau^2(U)]$.

References

1. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. *Statistics in Medicine*. 2014; 33(13):2297–2340. [PubMed: 24599889]
2. Stuart EA. Matching methods for causal inference: A review and a look forward. *Statistical Science*. 2010; 25(1):1. [PubMed: 20871802]

3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70(1):41–55.
4. Nason M, Follmann D. Design and analysis of crossover trials for absorbing binary endpoints. *Biometrics*. 2010; 66(3):958–965. [PubMed: 19930189]
5. Griffin BA, Lagakos S. Analysis of failure time data arising from studies with alternating treatment schedules. *Journal of the American Statistical Association*. 2006; 101(474):510–520.
6. Cook RJ, Wei W, Yi GY. Robust tests for treatment effects based on censored recurrent event data observed over multiple periods. *Biometrics*. 2005; 61(3):692–701. [PubMed: 16135020]
7. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *American journal of epidemiology*. 1991; 133(2):144–153. [PubMed: 1985444]
8. Lumley T, Levy D. Bias in the case–crossover design: implications for studies of air pollution. *Environmetrics*. 2000; 11(6):689–704.
9. Redelmeier DA, Tibshirani RJ. Interpretation and bias in case-crossover studies. *Journal of Clinical Epidemiology*. 1997; 50(11):1281–1287. [PubMed: 9393384]
10. Navidi W. Bidirectional case-crossover designs for exposures with time trends. *Biometrics*. 1998; 54(2):596–605. [PubMed: 9629646]
11. Luo X, Sorock GS. Analysis of recurrent event data under the case-crossover design with applications to elderly falls. *Statistics in Medicine*. 2008; 27(15):2890–2901. [PubMed: 18095315]
12. Farrington C. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics*. 1995:228–235. [PubMed: 7766778]
13. Whitaker HJ, Paddy Farrington C, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. *Statistics in Medicine*. 2006; 25(10):1768–1797. [PubMed: 16220518]
14. Huang CY, Qin J, Wang MC. Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics*. 2010; 66(1):39–49. [PubMed: 19459839]
15. Kalbfleisch JD. Likelihood methods and nonparametric tests. *Journal of the American Statistical Association*. 1978; 73(361):167–170.
16. Brahmadathan, JJ.; Abraham, K.; Morens, V., et al. Streptococcal pharyngitis in school children in southern india: A prospective natural history study, 2016. Working manuscript
17. Levin ML. The occurrence of lung cancer in man. *Acta-Unio Internationalis Contra Cancrum*. 1952; 9(3):531–541. [PubMed: 13124110]
18. Greenland S. Relation of probability of causation to relative risk and doubling dose: a methodologic error that has become a social problem. *American Journal of Public Health*. 1999; 89(8):1166–1169. [PubMed: 10432900]
19. Little RJ. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*. 1993; 88(421):125–134.
20. Lawless JF, Nadeau C. Some simple robust methods for the analysis of recurrent events. *Technometrics*. 1995; 37(2):158–168.
21. Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*. 1988; 44(1):175–188.
22. Follmann D, Wu M. An approximate generalized linear model with random effects for informative missing data. *Biometrics*. 1995; 51(1):151–168. [PubMed: 7766771]
23. Okeke IN, Lamikanra A, Edelman R. Socioeconomic and behavioral factors leading to acquired bacterial resistance to antibiotics in developing countries. *Emerging Infectious Diseases*. 1999; 5(1):18. [PubMed: 10081668]
24. Levy, SB. The antibiotic paradox. Da Capo Press; 2002.

U – quality of immune reaction to GAS exposures

Figure 1.

Schematic of how a time-constant unmeasured confounder U , a measure of the immune system's response to GAS exposure, impacts periods of colonization and episodes of pharyngitis.

Figure 2.

The presence of Group A *Streptococcus* in the throat for 22 selected children during the study. Monthly carriage visits are denoted by an open circle. Any weekly positive pharyngitis visit during the month is denoted by a closed circle. Green circles denote GAS negative visits, while red circles denote GAS positive visits.

Table 1

The four kinds of individuals and their proportions in the population during a small interval of time $[t, t + \Delta t)$. Different outcomes are observed depending on the exposure status, Z .

Label	Proportion	Event if $Z = 0$ $Y(0)$	Event if $Z = 1$ $Y(1)$
Immune	θ_{00}	0	0
Exposure caused disease	θ_{01}	0	1
Exposure prevented disease	θ_{10}	1	0
Doomed	θ_{11}	1	1

Table 2

Performance of different estimates of the probability of causation.

Simulation Settings						Probability of Causation					
β_2	β_3	σ_e	σ_v	$\overline{N}_+$	$\hat{p}_+$	Naive	Bin	Mult	Pair	$\mathcal{Z}$	Omni
0	0	1.0	1.0	201	0.512	0.633	0.631	0.632	0.633	0.631	0.630
						0.019	0.029	0.030	0.031	0.030	0.021
						0.001	-0.001	0.000	0.001	-0.001	-0.002
0	0	0.5	0.5	168	0.451	0.019	0.029	0.030	0.031	0.030	0.021
						0.634	0.634	0.636	0.635	0.634	0.629
						0.024	0.031	0.032	0.032	0.046	0.025
1	0	1.0	1.0	431	0.499	0.002	0.002	0.004	0.003	0.002	-0.003
						0.024	0.031	0.032	0.032	0.046	0.025
						0.807	0.630	0.630	0.631	0.632	0.635
1	0	0.5	0.5	211	0.449	0.009	0.014	0.016	0.017	0.026	0.025
						0.175	-0.002	-0.002	-0.001	0.000	0.003
						0.320	0.014	0.016	0.017	0.026	0.025
1	0	1.0	1.0	265	0.521	0.698	0.634	0.635	0.634	0.576	0.628
						0.015	0.024	0.026	0.026	0.067	0.034
						0.066	0.002	0.003	0.002	-0.056	-0.004
1	-1	0.5	0.5	187	0.456	0.059	0.024	0.026	0.026	0.098	0.034
						0.487	0.244	0.249	0.230	0.384	0.381
						0.059	0.200	0.190	0.170	0.100	0.098
1	-1	1.0	1.0	201	0.512	0.093	-0.150	-0.145	-0.164	-0.010	-0.013
						0.145	0.425	0.400	0.439	0.101	0.100
						0.598	0.536	0.539	0.528	0.582	0.580
1	0	0.5	0.5	168	0.451	0.030	0.059	0.058	0.062	0.049	0.032
						0.015	-0.047	-0.044	-0.055	-0.001	-0.003
						0.032	0.081	0.077	0.092	0.049	0.032

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

NOTE: The proportional rate function is given by $\exp(0 + Z(\theta)1 + U\beta_2 + UZ(\theta)\beta_3)$ with $U \stackrel{D}{=} N\left(-2, \sigma_U^2\right)$. Each set of four rows provides medians, robust variance ($\times 10$), median bias, and robust MSE ($\times 10$) for the different estimators over 1000 simulated studies, each with 200 subjects. Columns 5 and 6 report the average number of events and proportion of patients with at least 1 event, averaged over the simulations.

Table 3

Performance of different estimates of the probability of causation.

Simulation Settings						Probability of Causation					
β_2	β_3	σ_e	σ_v	$\overline{N_+}$	$\hat{p}_+$	Naive	Bin	Mult	Pair	$\mathcal{I}$	Omni
0	0	1.0	1.0	220	0.533	0.631	0.200	0.635	0.634	0.627	0.629
						0.041	0.160	0.083	0.087	0.120	0.061
						-0.001	-0.432	0.003	0.002	-0.005	-0.003
0	0	0.5	0.5	184	0.468	0.041	2.000	0.083	0.087	0.120	0.061
						0.631	0.183	0.634	0.636	0.636	0.625
						0.087	0.180	0.120	0.110	0.190	0.110
1	0	1.0	1.0	467	0.510	-0.001	-0.449	0.002	0.004	0.004	-0.007
						0.087	2.200	0.120	0.110	0.190	0.110
						0.886	0.201	0.620	0.622	0.644	0.634
1	0	0.5	0.5	232	0.467	0.008	0.070	0.043	0.041	0.079	0.045
						0.254	-0.431	-0.012	-0.010	0.012	0.002
						0.650	1.900	0.044	0.042	0.080	0.045
1	0	0.5	0.5	232	0.467	0.771	0.187	0.639	0.641	0.578	0.633
						0.033	0.140	0.081	0.083	0.250	0.084
						0.139	-0.445	0.007	0.009	-0.054	0.001
1	-1	1.0	1.0	210	0.472	0.230	2.100	0.081	0.084	0.280	0.084
						0.555	-0.860	-0.016	-0.015	0.377	0.383
						0.066	1.300	0.750	0.720	0.260	0.160
1	-1	0.5	0.5	213	0.472	0.161	-1.254	-0.410	-0.409	-0.017	-0.011
						0.325	17.000	2.430	2.390	0.263	0.161
						0.627	-0.056	0.471	0.477	0.567	0.575
1	-1	1.0	1.0	213	0.472	0.076	0.310	0.210	0.210	0.240	0.110
						0.044	-0.639	-0.112	-0.106	-0.016	-0.008
						0.095	4.390	0.335	0.322	0.243	0.110

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

NOTE: The proportional rate function is given by $\exp(A_k + Z\theta_1 + U\beta_2 + UZ\theta\beta_3)$ for $k = 1, \dots, 5$ with $U \stackrel{D}{=} N\left(-2, \sigma_U^2\right)$. Each set of four rows provides medians, robust variance ($\times 10$), median bias, and robust MSE ($\times 10$) for the different estimators over 1000 simulated studies, each with 200 subjects. Columns 5 and 6 report the average number of events and proportion of patients with at least 1 event, averaged over the simulations.

Table 4

Episodes of pharyngitis (sore throat) and carriage visits classified by the presence of Group A *Streptococcus* in the throat.

	Carriage Data Monthly Visit	Event Data # Pharyngitis Events
No Group A Strep	2504	530
Group A Strep	323	110
Total	2827	640