Title
Inference of Naturally Acquired Immunity Using a Self-matched Negative-Control Design.

Permalink
https://escholarship.org/uc/item/96x8t52k

Journal
Epidemiology (Cambridge, Mass.), 32(2)

ISSN
1044-3983

Authors
Northrup, Graham R
Qian, Lei
Bruxvoort, Katia
et al.

Publication Date
2021-03-01

DOI
10.1097/ede.00000000000001305

Peer reviewed
Abstract: Host adaptive immune responses may protect against infection or disease when a pathogen is repeatedly encountered. The hazard ratio of infection or disease, given previous infection, is typically sought to estimate the strength of protective immunity. However, variation in individual exposure or susceptibility to infection may introduce frailty bias, whereby a tendency for infections to recur among individuals with greater risk confounds the causal association between previous infection and susceptibility. We introduce a self-matched “case-only” inference method to control for unmeasured individual heterogeneity, making use of negative-control endpoints not attributable to the pathogen of interest. To control for confounding, this method compares event times for endpoints due to the pathogen of interest and negative-control endpoints during counterfactual risk periods, defined according to individuals’ infection history. We derive a standard Mantel-Haenszel (matched) odds ratio conveying the effect of prior infection on time to recurrence. We compare performance of this approach to several proportional hazards modeling frameworks and estimate statistical power of the proposed strategy under various conditions. In an example application, we use the proposed method to reestimate naturally acquired protection against rotavirus gastroenteritis using data from previously published cohort studies. This self-matched negative-control design may present a flexible alternative to existing approaches for analyzing naturally acquired immunity, as well as other exposures affecting the distribution of recurrent event times.

Key Words: natural immunity, frailty, negative control, hazard, self-matched, Mantel-Haenszel

Epidemiology 2021;32: 168–178

Host adaptive immune responses often protect against infection or disease when a pathogen is repeatedly encountered. Vaccines aim to exploit this protection by exposing hosts to an attenuated infection or to immunizing sub-units of a pathogen. Evidence of protective naturally acquired immunity thus provides a strong rationale for vaccine development.1 Quantitative estimates of the strength of naturally acquired protection also inform the interpretation of epidemiologic data, for instance providing a baseline against which vaccine performance can be evaluated.2 These estimates are further sought to parameterize mathematical models of pathogen transmission.3

Naturally acquired immunity is often estimated via the hazard ratio of infection or disease, comparing counterfactual periods in the presence and absence of prior infection.4–10 Thus, inference centers on the distribution of recurrent event times. Unmeasured heterogeneity in individuals’ hazards of infection or disease presents a challenge in such analyses, originally termed a problem of “varying liabilities” by Greenwood and Yule11 and subsequently addressed as “accident-proneness”12 or “frailty.”13 The tendency for events to recur among certain individuals must be accounted for in statistical analyses14, recurrence of infection or disease among individuals with the greatest susceptibility or exposure to a pathogen, irrespective of previous infection, may bias estimates of naturally acquired protection.15

This consideration may have relevance to several diseases against which immune responses generate imperfect protection. Tuberculosis presents a notable example, where despite evidence of protective cell-mediated and humoral
immunity.\textsuperscript{16} Several epidemiologic studies have reported higher rates of new-onset infection or disease among persons previously treated successfully for active tuberculosis, as compared to those without history of tuberculosis.\textsuperscript{17–20} Similar conflict about the consequences of prior infection has arisen in epidemiologic studies of gonorrhea.\textsuperscript{21,22} In recent analyses of a multi-site pediatric cohort study addressing enteric disease, previous infection predicted higher rates of recurrent infection or disease associated with several pathogens, including \textit{Shigella spp.}, \textit{Campylobacter spp.}, and various diarrheagenic \textit{Escherichia coli} strains.\textsuperscript{23} Evidence supporting the feasibility of protective vaccines against many of these pathogens suggests a need to revisit the impacts of naturally acquired immunity.\textsuperscript{24–26} Similar causal inference challenges arise in the relationship between chronic inflammation and repeated infection in conditions such as cystic fibrosis,\textsuperscript{27,28} otitis media,\textsuperscript{29,30} and environmental enteric dysfunction.\textsuperscript{31}

Formalizing unmeasured heterogeneity as confounding factors potential strategies to identify naturally acquired protection. Terming $Y_i$ and $Y_2$ as primary and recurrent infection or disease outcomes, respectively, and $U$ as the constellation of unmeasured individual factors influencing exposure or susceptibility to a pathogen of interest, a directed acyclic graph (Figure 1) reveals that $Y_1 \leftarrow U \rightarrow Y_2$ may introduce bias into the estimation of $Y_1 \rightarrow Y_2$, the effect of primary infection on recurrence. Conditioning on unmeasured individual factors by comparing observations during counterfactual risk periods from the same individual ($Y_1 \leftarrow U \rightarrow Y_2$) permits unbiased inference of the effect of $Y_1$. This intuition provides the basis for numerous self-matched designs (e.g., case-crossover, case-time control, and self-controlled case series), which have garnered increasing interest in epidemiology.\textsuperscript{32}

In this article, we present an adaptation of these methods harnessing data from “negative control” events to permit causal inference in the presence of heterogeneous individual frailty. We derive a matched (Mantel-Haenszel) odds ratio ($OR_{MH}$)\textsuperscript{33,34} estimator for the hazard ratio of infection or disease, given previous infection. We conduct simulations to compare this approach against alternative methods based on proportional hazards models common in the analysis of longitudinal data and to assess statistical power under varying conditions. Last, we use the proposed method to reassess protective effects of rotavirus infection in data from previously published birth-cohort studies.\textsuperscript{4,5}

\section*{APPRAOCH}

\subsection*{Self-matched Negative-Control Design}

Consider an outcome such as acquisition of a pathogen of interest, or onset of disease due to this pathogen (Table 1). The proposed design only includes individuals who experience recurrent episodes of this outcome of interest (case-only). Define $Y_i$ and $X_i$ as variables indicating outcome and exposure status for individual $i$ at each observation, with $Y_i = 1$ indicating infection or disease with the pathogen of interest and $Y_i = 0$ indicating a negative-control outcome. Consideration of negative-control observations is of interest for studies involving event-based data capture (e.g., episodes of acute illness) and provides a basis for competing risks estimation frameworks as we detail below. Last, let $X_i = 1$ indicate an individual has previously experienced infection with the pathogen of interest, and let $X_i = 0$ indicate the individual has no history of infection with the pathogen of interest.

Define $A_i$ to $D_i$ as random variables indicating event times for observations of $Y_i = 1$ and $Y_i = 0$, conditioned on $X_i$ (Table 2). $A_i$ and $B_i$ are the time to first occurrence of the outcome of interest and the negative-control outcome, respectively, for an individual with no history of infection ($X_i = 0$). $C_i$ and $D_i$ are the time to the first occurrence of the outcome of interest and the negative-control outcome, respectively, following infection with the pathogen of interest (such that $X_i = 1$; Figure 2). Here, we note that $B_i$ and $D_i$ are censored if $A_i < B_i$ and $C_i < D_i$, respectively.

\subsection*{Event Time Distributions}

Define the hazard for the “positive” outcomes owing to infection with the pathogen of interest ($P$) for individual $i$ as $\lambda_{Pi}$, and define $\theta$ as the hazard ratio of this outcome given

\begin{table}[ht]
\centering
\begin{tabular}{|l|l|}
\hline
\textbf{Parameter} & \textbf{Definition} \\
\hline
$\lambda_{Pi}$ & Rate at which individual $i$ experiences a prespecified clinical endpoint due to the pathogen of interest (“outcome of interest”), in the absence of naturally acquired immunity \\
$\lambda_{Ni}$ & Rate at which individual $i$ experiences a negative-control outcome \\
$\theta$ & Hazard ratio for the outcome of interest, owing to naturally acquired protection \\
$\beta_P(1)$ and $\beta_P(0)$ & Hazard ratio for the outcome of interest during the period after primary infection, relative to the period before primary infection, for individual $i$, due to all (confounding) factors other than naturally acquired protection \\
$\beta_N(1)$ and $\beta_N(0)$ & Hazard ratio for the negative-control outcome during the period after primary infection, relative to the period before primary infection, for individual $i$ \\
\hline
\end{tabular}
\caption{Parameters and Definitions}
\end{table}

\begin{figure}
\centering
\begin{tikzpicture}
\node at (0,0) (A) {$Y_1$};
\node at (1,0) (B) {$Y_2$};
\draw[->] (A) -- (B);
\end{tikzpicture}
\caption{Directed acyclic graph addressing unmeasured confounding. We illustrate a causal framework wherein the effect of previous infection on time to subsequent infection ($Y_1 \rightarrow Y_2$) is of interest for analysis. One or more unmeasured confounding factors ($U$) creates a backdoor path (A) which can be blocked by conditioning on $U$ (B).}
\end{figure}

© 2020 The Author(s). Published by Wolters Kluwer Health, Inc.
previous infection (Table 1). Assuming infectious exposures occur independently and continuously during the follow-up period, conditioning on each individual’s unique hazard, event times are exponentially distributed at the individual level. We note this circumstance gives rise to a nonexponential event time distribution at the population level, with variance augmented (relative to the pure exponential case) by heterogeneity in individual frailty. The probability of the outcome by time \( t \), for a previously uninfected individual, is \( 1 - \exp(-\lambda_P t) \), while the probability of the outcome by time \( t \), had the same individual counterfactually been previously infected, is \( 1 - \exp(-\theta \lambda_P t) \). We discuss alternative event time distributions in a later section.

Consider that data are collected from each individual for endpoints besides the primary outcome of interest. Among these, suppose a negative-control outcome \((N)\) occurs at a rate \( \lambda_N \), for individual \( i \). This rate should be unaffected by individuals’ prior exposure to the pathogen of interest, according to the definition of a negative control in this context. 

Under the same assumptions, the probability of experiencing the negative-control outcome by time \( t \), for individual \( i \), is 
\[
1 - \exp\left(\frac{-\lambda_N t}{\lambda_P + \lambda_N}\right)
\]

**TABLE 2.** Contingency Table for Event Time Distributions, Given Prior Infection

| Exposure Status | Outcome of Interest | Negative Control Outcome |
|-----------------|---------------------|--------------------------|
| Previously uninfected | \( X_i = 0 \) \( A_i \sim \text{Exp}(\lambda_P) \) | \( B_i \sim \text{Exp}(\lambda_N) \) |
| Previously infected | \( X_i = 1 \) \( C_i \sim \text{Exp}(\theta \lambda_P) \) | \( D_i \sim \text{Exp}(\lambda_N) \) |

**FIGURE 2.** Schematic presentation of potential outcomes. We illustrate potential outcomes in terms of the sequence of events \( A_i-D_i \) for a given individual. In cases 1 and 2, we observe \( A_i < B_i \) (truncating observation of a negative-control event although \( X_i = 0 \)). In cases 3 and 4, we observe \( B_i < A_i \), with the negative-control outcome preceding infection with the pathogen of interest. We illustrate the corresponding potential outcomes for \( C_i \) and \( D_i \), when \( X_i = 1 \), in the right-hand side of the figure.

**Estimating the Effect of Naturally Acquired Immunity**

For an individual with no history of previous infection, consider the outcome of interest and the negative-control outcome to be competing risks. The observations \( E_i \) to \( H_i \) may be defined to indicate the relative ordering of event times \( A_i \) to \( D_i \) (Table 3), for each individual \( i \). Specifically, take \( E_i = A_i \leq B_i \) and \( G_i = C_i \leq D_i \) to indicate the outcome of interest precedes the negative-control outcome during the periods with \( X_i = 0 \) and \( X_i = 1 \), respectively. Define \( F_i = B_i < A_i \) and \( H_i = D_i < C_i \) as complements to \( E_i \) and \( G_i \). By the memoryless property of the exponential distribution, we may start the clock for event times \( A_i \) and \( B_i \) at cohort entry or at any alternative milestone of interest (e.g., birth or exposure onset); for \( C_i \) and \( D_i \), we consider time from the most recent infection, although any subsequent milestone is likewise appropriate.

As formulated in Appendices A and B, we have

\[
\text{Pr}(E_i) = \text{Pr}(A_i \leq B_i) = \frac{\lambda_P}{\lambda_P + \lambda_N}
\]


is the regularized incomplete beta function is unaffected. Formally, we
\( EA B \)
\( t \)

\( \sum_{i} \theta \eta _{i} \theta _{N} \theta _{i} \theta _{N} \)

\( \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)

\( \theta \lambda _{N} (t) \)
the distributions. We illustrate potential bias under these alternative parameterizations in eFigures 2 and 3 (http://links.lww.com/EDE/B749).

**COMPARISON TO COHORT DESIGN USING PROPORTIONAL HAZARDS ANALYSIS**

**Simulation Study**

We conducted a simulation study across various underlying distributions of $\lambda_{pi}$ and $\lambda_{Ni}$ to test for bias of point estimates under the proposed approach and under alternative methods often used in the analysis of cohort study data (without consideration of negative controls). As comparisons, we considered several proportional hazards models, which could be applied to time-to-event data for recurrent observations of the outcome of interest. We considered four approaches to control for differences in hazards among individuals:

1. “Naive” proportional hazards model without inclusion of additional terms to account for differences in event times among individuals. We define the resulting hazard ratio $\hat{\theta}_{\text{Naive}}$.
2. Proportional hazards model accounting for variation in individual frailty via “random effects.” Fitting this model yields $\hat{\theta}_{RE}$ for the effect of previous infection, along with $\hat{\sigma}^2$ representing the estimated variance in (log) individual-specific event rates, assumed to represent independent draws from a Normal distribution with mean 0.37,38
3. Proportional hazards model including Gamma-distributed frailty terms.13 Fitting this model yields $\hat{\theta}_{\text{Fraily}}$ for the effect of previous infection, along with the parameters of the underlying Gamma distribution describing individual-specific frailties.
4. Proportional hazards model with “fixed effects” for individual subjects. Fitting this model yields $\hat{\theta}_{FE}$ for the effect of previous infection and estimates subject-specific rates of infection (via. individual-specific intercepts), which have no preassumed distribution.

We defined $\hat{\theta}_{MH} = OR_{MH}$ for the proposed analysis strategy of a self-matched, negative-control design and considered various distributions for $\lambda_{pi}$:

1. Truncated Normal distribution (with a prespecified lower bound at $a = 0$);
2. Truncated Cauchy distribution (with a prespecified lower bound at $a = 0$);
3. Uniform distribution;
4. Gamma distribution;
5. Mixtures of Gamma distributions.

We considered multiple parameterizations of each of these distributions (Table 4), holding the mean rate (or location parameter of the Cauchy distribution) constant at one infection per year across all simulations to determine effects of inter-individual heterogeneity on estimates of $\theta$. We illustrate the distributions in Figure 3. Considering cohorts of 500 individuals, we drew $\lambda_{pi}$ values at random and sampled exponentially distributed event times of first and second infections for each individual, truncating observations at five years. We repeated simulations 500 times for each $\theta \in \{0.01, 0.02, \ldots, 0.99\}$, drawing $\lambda_{pi}$ values independently for each simulation. We used the simulated datasets to estimate $\hat{\theta}_{\text{Naive}}, \hat{\theta}_{RE}, \hat{\theta}_{FE},$ and $\hat{\theta}_{\text{Fraily}},$ taking the average of estimates obtained across all 500 iterations to obtain a single point estimate for each parameterization.

To compute $\hat{\theta}_{MH}$, we drew hazards ($\lambda_{pi}$) and event times for negative-control observations for each subject, assuming event times were exponentially distributed with respect to the sampled, individual-specific rate parameters. To standardize comparisons of $\hat{\theta}_{MH}$ under differing distributions of $\lambda_{pi}$, we defined $\lambda_{Ni} = 1 \forall i$ under each simulation.

To investigate how the different modeling frameworks performed in capturing the distribution of individual-specific hazards, we saved estimates of individual-specific fixed effects, random effects, and frailties alongside estimates of $\theta$.

We fitted a single density kernel to the distribution of individual-specific estimates across 10 simulated cohorts for each true value of $\theta$ and under sampling of $\lambda_{pi}$.

**RESULTS**

We plot distributions and estimates under each approach in Figure 3. The naive hazards ratio tended to overestimate of the causal effect $\theta$, leading to under-estimation of the degree of protection ($1 - \theta$). Bias was minimized as $\theta$ approached zero, consistent with a scenario of strong protective immunity.

Values of $\hat{\theta}_{\text{Naive}}$ often exceeded 1 in scenarios where $\theta < 1$; in practice, such an estimate would lead to inference that prior infection increases susceptibility to infection or disease attributable to the pathogen of interest, when in fact prior infection is protective. For all distributions considered, bias in $\hat{\theta}_{\text{Naive}}$ was greatest under parameterizations yielding the highest between-individual variance in $\lambda_{pi}$.

Alternative methods performed variably under the differing conditions (Figure 3). Lower degrees of bias were evident in $\hat{\theta}_{MH}$ as compared to estimates generated under the other methods assessed. Gamma frailty models and random effects models tended to yield less-biased estimates of $\theta$ than $\hat{\theta}_{\text{Naive}}$. However, the same direction of bias (resulting in under-estimation of the reduction in susceptibility, or $\hat{\theta} > \theta$) was evident with all three of these approaches. Bias was worst when $\lambda_{pi}$ values were drawn from Gamma or Gamma mixture distributions and tended to increase under distributions with greater variance in $\lambda_{pi}$, or greater irregularity in the case of Gamma mixture distributions. In contrast, fixed-effects models estimating multipliers on hazards for each individual tended to under-estimate $\theta$ under most distributions of $\lambda_{pi}$, although both $\hat{\theta}_{FE} > \theta$ and $\hat{\theta}_{FE} < \theta$ were apparent in simulations using the truncated Cauchy distribution for $\lambda_{pi}$. For the truncated Normal distribution, bias in $\hat{\theta}_{FE}$ decreased with greater variance in $\lambda_{pi}$.
whereas for the Uniform, Gamma, and Gamma mixture distributions, bias increased with greater variance in \( \lambda_{pi} \).

Biased estimation of \( \theta \) occurred in connection with a failure to accurately recover the underlying individual-specific frailty distributions. For each modeling approach, the extent of this misspecification in individual frailties varied over values of \( \theta \) and distributions of \( \lambda_{pi} \) (eFigures 4 to 6; http://links.lww.com/EDE/B749).

### SAMPLE SIZE CONSIDERATIONS

#### Simulation Study

To inform applications of the proposed method, we next assessed statistical power under differing conditions. A test statistic \( S_{MH} \) has previously been identified for \( OR_{MH} \) under the null hypothesis of no difference in risk given exposure.\(^{39}\) For the contingency structure (Table 3) formulated from the terms \( E_i \) to \( H_i \), this statistic can be written generally as

\[
S_{MH} = \frac{\sum_i^N (E_i - (E_i + F_i)(E_i + G_i))^2}{\sum_i^N (E_i + F_i)(E_i + G_i)(F_i + H_i)(G_i + H_i)}.
\]  

which can then be simplified according to \( E_i + F_i = 1 \), \( G_i + H_i = 1 \), and \( F_i + H_i = 2 - E_i - G_i \). Thus,

\[
S_{MH} = \frac{\sum_i^N (E_i - G_i)^2}{\sum_i^N (E_i + G_i)(2 - E_i - G_i)},
\]

which is expected to follow a \( \chi^2 \) distribution with one degree of freedom under the null hypothesis. We calculated values of \( \tilde{S}_{MH} \) obtained for cohorts of varying sizes under differing parameterizations of \( \theta, \lambda_{pi} \), and \( \lambda_{NH} \). For values of \( \theta \in \{0.1, 0.2, \ldots, 0.9\} \), we sampled individual event times from \( A_i \) to \( D_i \) for a population of 100,000 individuals whom we subsequently partitioned (without replacement) into 2,000 hypothetical study cohorts each of sizes from \( N=25 \) to \( N=1,000 \). For these analyses, we considered \( \lambda_{pi} \) values drawn from truncated Normal, Gamma, and Gamma mixture distributions, under the parameterizations of each of these distributions with greatest and least variance listed in Table 4. We determined statistical power via the proportion of simulated cohorts for which the upper bound of a 95% confidence interval around \( OR_{MH} \) would be expected to correctly exclude the null value, that is, \( \Pr(S_{MH} > \tilde{S}_{MH}^{0.95} = 5.02) \).

### Table 4. Event Rate Distributions Applied to Simulation Study

| Distribution       | Parameters               | Parameterizations* |
|--------------------|--------------------------|--------------------|
| Truncated Normal   | Mean \( \mu \)           | \( \mu = 1 \)       |
|                    | Variance \( \sigma^2 \)   | \( \sigma = 1/4 \)  |
|                    | Lower bound a            | a = 0              |
|                    | Upper bound a            | b = \( \infty \)    |
| Truncated Cauchy   | Location \( x_0 \)        | \( x_0 = 1 \)       |
|                    | Scale \( \gamma \)        | \( \gamma = 1/8 \)  |
|                    | Lower bound a            | a = 0              |
|                    | Upper bound a            | b = \( \infty \)    |
| Uniform            | Lower bound a            | a = 7/8            |
|                    | Upper bound a            | b = 9/8            |
| Gamma              | Shape k                  | k = 8              |
|                    | Scale \( \theta \)        | k = 1/8            |
| Gamma mixture (i)  | Shapes \( k_1, k_2 \)     | \( k_1 = 1/8 \)     |
|                    | Scale \( \theta_1, \theta_2 \) | \( \theta_1 = 1 \) |
|                    | Weight \( \omega \)        | \( \omega = 0.5 \)  |
| Gamma mixture (ii) | Shapes \( k_1, k_2 \)     | \( k_1 = 1/2 \)     |
|                    | Scale \( \theta_1, \theta_2 \) | \( \theta_1 = 1 \) |
|                    | Weight \( \omega \)        | \( \omega = 0.5 \)  |

*Parameterizations are listed in order of increasing variance from I to V.

The weight parameter of the Gamma mixture distribution indicates the proportion of individuals whose rates are parameterized according to \( k_1 \) and \( \theta_1 \); the proportion with rates parameterized according to \( k_2, \theta_2 \) is \( 1 - \omega \).
FIGURE 3. Simulated distributions and hazard ratio estimates under “naive” inference approaches and under the proposed approach of self-matched inference with negative controls. Panels are organized to present, in each row, the assumed distribution (column 1), the estimate $\hat{\theta}_{\text{Naive}}$ based on a Cox proportional hazards model without any correction for inter-individual heterogeneity (column 2), proportional hazards models employing various frailty frameworks (columns 3–5), and the estimate $\hat{\theta}_{\text{MH}}$ based on the proposed approach (column 6). One-to-one lines plotted in gray in columns 2–6 indicate where estimates would recover the true value, that is, $\hat{\theta} = \theta$. Horizontal gray lines plotted at $\hat{\theta} = 1$ indicate where estimates exceed 1, indicating directionally misspecified estimates of the causal effect of interest. Values are plotted on a red-to-blue color ramp corresponding to the parameterizations I-V, respectively, in order of least (I; red) to greatest (V; blue) variance as detailed in Table 4. (A) Truncated normal distribution; (B) truncated Cauchy distribution; (C) uniform distribution; (D) Gamma distribution; (E) mixture of Gamma distributions (i) with means at 0.125 and 1.875; and (F) mixture of Gamma distributions (ii) with means at 0.5 and 1.5.
To assess how correlation between $\lambda_{Ni}$ and $\lambda_{Pi}$ could affect the statistical power of estimates, we conducted simulations under two sets of assumptions. Under the first, we considered $\lambda_{Ni} = 1 \forall i$ (equal to the expected value of $\lambda_{Pi}$ under all parameterizations), so that the correlation would be zero; under the second, we defined $\lambda_{Ni} = \lambda_{Pi}$, under the assumption that individuals with greater risk of the outcome of interest would also experience higher incidence of the negative-control condition. These conditions provided upper and lower bounds on power, corresponding to assumptions of no correlation and perfect correlation between $\lambda_{Ni}$ and $\lambda_{Pi}$, respectively.

RESULTS

We present results of the power analyses in Figure 4. Analyses with as few as 50 subjects had roughly 80% power or greater to estimate $\theta = 0.1$ (corresponding to 90% protection) under all conditions explored; analyses with 500 subjects had 80% power or greater to estimate $\theta \leq 0.5$ (corresponding to 50% protection or greater) under all conditions. No scenarios revealed 80% or greater power for estimation of $\theta \geq 0.8$ (corresponding to less than 20% protection), even with 1,000 subjects; statistical power for estimation of $\theta = 0.9$ was 10% or lower under nearly all conditions explored. This limitation makes our proposed method potentially inefficient for scenarios where immunity is very weak; generally, the use of stochastic negative-control observations for inference diminishes statistical power in comparison to approaches that do not require such data.

For simulations with $\lambda_{Ni} \perp \lambda_{Pi}$, statistical power was weaker under parameterizations resulting in greater variance in $\lambda_{Pi}$. In contrast, for simulations with $\lambda_{Ni} = \lambda_{Pi}$, differences in statistical power were less strongly apparent with increasing variance in $\lambda_{Pi}$. Taken together, these findings suggest statistical power is maximized when negative-control endpoints are chosen which tend to occur more commonly among individuals who are at greatest risk of the outcome of interest.

APPLICATION TO ROTAVIRUS BIRTH-COHORT DATA

Last, we applied the proposed method to real-world data collected in two birth-cohort studies of rotavirus infection and disease among 200 children in Mexico City, Mexico, and 373

![FIGURE 4](image-url). Statistical power for simulated analyses using the proposed approach of self-matched inference via negative controls. Each panel presents the statistical power for rejecting the null hypothesis with two-sided $P < 0.05$ under varying conditions. Lines plotted in red to blue correspond to decreasing values of $\theta$: 0.9 (red), 0.8, 0.7, ..., 0.1 (blue), corresponding to increasing protection from 10% to 90%. Plots are presented in groups of four panels, each corresponding to analyses with values drawn from the following distributions: (A) Truncated Normal distribution; (B) Gamma distribution; (C) Mixture of Gamma distributions with means at 0.125 and 1.875 (as detailed in Table 4). Panels in the top row (A.1, A.2, B.1, B.2, C.1, C.2) represent analyses in which no correlation is assumed between rates of the outcome of interest and negative control outcome ($\lambda_{Ni} = 1 \forall i$). Panels in the bottom row (A.3, A.4, B.3, B.4, C.3, C.4) represent analyses in which the correlation between rates of the outcome of interest and the negative control outcome are maximized ($\lambda_{Ni} = \lambda_{Pi}$). Within each grouping, panels on the left-hand side (A.1, A.3, B.1, B.3, C.1, C.3) correspond to distributions with the least variance in individual rates of the outcome of interest ($\lambda_{Pi}$; i.e., parameterization I in Table 4). Panels on the right-hand side within each grouping (A.2, A.4, B.2, B.4, C.2, C.4) correspond to distributions with the greatest variance in individual rates of the outcome of interest ($\lambda_{Pi}$; i.e., parameterization V in Table 4).
children in Vellore, India. These datasets have been described extensively in primary study publications\(^4\,5\) and subsequent reanalyses\(^{15,40}\); we present details in the Supplemental Digital Content (http://links.lww.com/EDE/B749).

Initial analyses of the datasets gave differing conclusions about the strength of protection against rotavirus gastroenteritis (RVGE). Based on Cox proportional hazards models that did not account for individual frailty, children in Mexico City were estimated to have experienced 77% (95% confidence interval = 60, 88), 83% (64–92), and 92% (44–99) lower rates of RVGE following one, two, and three previous infections, respectively, as compared to zero infections.\(^4\) In contrast, children in Vellore, where the rate of rotavirus acquisition was higher, were estimated to have experienced 43% (24–56%), 71% (59–80%), and 81% (69–88%) lower rates of RVGE after one, two, and three previous infections, as compared to zero infections.\(^5\) Subsequent analyses of the datasets revealed substantial variation in rates of rotavirus infection and risk of RVGE among individual children, as well as a potential for confounding due to declining risk of RVGE when infections were acquired at older ages, irrespective of previous infection.\(^{40}\) In contrast, model-based analyses accounting for the independent effects of age and previous infection on children’s susceptibility to RVGE estimated that children experienced 33% (23–41%), 50% (42–57%), and 64% (55–70%) lower rates of RVGE after one, two, and three previous infections, respectively, as compared to zero infections.\(^{15}\)

We used the proposed self-matched negative-control design to reestimate naturally acquired protection against RVGE in the cohort datasets. Here, RVGE episodes (acute, new-onset diarrhea with rotavirus detected in the stool) were the outcome of interest and acute, new-onset diarrhea episodes without rotavirus detection were the negative controls. Our analyses did not adjust for age (owing to the limited number of children with 2 or 3 infections observed) or seasonality (due to year-round rotavirus transmission in these high-incidence settings). We compared the times of RVGE and rotavirus-negative diarrhea episodes from each child beginning from birth and thereafter following detection of the first, second, and third rotavirus infection, generating confidence intervals via resampling of individual children. This yielded estimates of 27% (–1–48%), 50% (13–73%), and 48% (0–77%) lower rates of RVGE following one, two, and three previous infections, as compared to zero infections (Figure 5). Notwithstanding lower statistical power for the proposed method, these estimates are in agreement with previous findings\(^{15}\) suggesting lower strength of naturally acquired protection than what was estimated in initial analyses of the studies.\(^{4,5}\)

**DISCUSSION**

We propose a novel self-matched negative-control method for estimating naturally acquired protection against recurrent infection or disease endpoints associated with a pathogen of interest. Analytically and via. simulation, we show this method recovers unbiased estimates under a range of conditions, including when individual hazards are drawn from highly irregular or skewed distributions. We find such scenarios may lead to bias under proportional hazards models with commonly used frailty estimation frameworks. Desirably, the proposed approach requires no parametric assumptions about individual-specific frailty. While our approach assumes exponentially distributed event times, this can be expected to hold provided infectious exposures occur continuously and independently, according to the individual-specific hazards. Beyond infectious disease natural history studies, this approach may have value for assessing the effects of other exposures on recurrent event times.

Our findings provide several practical insights for real-world longitudinal cohort studies. Collecting data on multiple endpoints affords the opportunity to leverage negative-control observations to support causal inference. For studies applying the proposed approach, negative-control
endpoints affected by the same risk factors or exposures as the outcome of interest are desirable both to reduce potential risks of confounding due to time-varying factors and to maximize statistical power based on the correlation between event rates for the outcome of interest and the negative-control outcome. “Test-negative” control conditions which resemble the outcome of interest, but are not attributable to the same pathogen, may provide a compelling choice, particularly if their occurrence is predicted by similar factors. For instance, shared risk factors are well-documented for rotavirus-positive and rotavirus-negative diarrhea. Considering respiratory illness, multiple etiologic viruses may share similar seasonal transmission patterns, routes of transmission via high-risk contact, and predictors of severe illness. For sexually transmitted infections, particular risk behaviors differing among individuals or over time could alter risk of any infection, rather than infection with the pathogen of interest alone. In the context of real-world cohort studies, test-negative control conditions that are clinically similar to the outcome of interest would likely result in a study visit or other recorded interaction with similar probability. This further supports consideration of inference methods making use of test-positive and test-negative occurrences of a particular clinical syndrome.

In summary, self-matched inference via negative controls may provide a flexible strategy to circumvent bias introduced by variation in individual frailty for analyses of naturally acquired immunity. Applications to other exposures affecting the distribution of recurrent event times merit consideration, given the possible limitations we identify in existing analysis frameworks.

REFERENCES

1. Lopman B, Kang G. In praise of birth cohorts: norovirus infection, disease, and immunity. Clin Infect Dis. 2014;58:492–494.
2. Pasetti ME, Levine MM. Insights from natural infection-derived immunity to cholera instruct vaccine efforts. Clin Vaccine Immunol. 2012;19:1707–1711.
3. Heisterbeek H, Anderson RM, Andreasen V, et al; Isaac Newton Institute IDD Collaboration. Modeling infectious disease dynamics in the complex landscape of global health. Science. 2015;347:aaa4339.
4. Velázquez FR, Matson DO, Calva JJ, et al. Rotavirus infection in infants as protection against subsequent infections. N Engl J Med. 1996;335:1022–1028.
5. Gladstone BP, Ramani S, Mukhopadhyay I, et al. Protective effect of natural rotavirus infection in an Indian birth cohort. N Engl J Med. 2011;365:337–346.
6. Rouhani S, Peñataro Yori P, Paredes Oortegui M, et al; Etiology, Risk Factors, and Interactions of Enteric Infections and Malnutrition and the Consequences for Child Health and Development Project (MAL-ED) Network Investigators. Norovirus infection and acquired immunity in 8 countries: results from the MAL-ED Study. N Engl J Med. 2012;367:1201–1210.
7. King AA, Ionides EL, Pascual M, Bouma MJ. Inapparent infections and cholera dynamics. Nature. 2008;454:877–880.
8. Andrews JR, Noubary F, Wilensky RP, Corda R, Losina E, Horsburgh CR. Risk of progression to active tuberculosis following reinfection with Mycobacterium tuberculosis. Clin Infect Dis. 2012;54:784–791.
9. Katzelnick LC, Greesh L, Halloran ME, et al. Antibody-dependent enhancement of severe dengue disease in humans. Science. 2017;358:929–932.
10. Rodriguez-Barruquer I, Arnaitewie E, Jagannathan P, et al. Quantifying heterogeneous malaria exposure and clinical protection in a cohort of Ugandan Children. J Infect Dis. 2016;214:1072–1080.
11. Greenwood M, Yule GU. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. J R Stat Soc. 1920;83:25–48.
12. Arbous AG, Kerrich JE. Accident statistics and the concept of accident-proneness. Biometrics. 1951;7:340–432.
13. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography. 1979;16:439–454.
14. Glynn RJ, Buring JE. Ways of measuring rates of recurrent events. BMJ. 1996;312:364–367.
15. Lewnard JA, Lopman BA, Parashar UD, et al. Heterogeneous susceptibility to rotavirus infection and gastroenteritis in two birth cohort studies: Parameter estimation and epidemiological implications. PLoS Comput Biol. 2019;15:e1007014.
16. Achkar JM, Casadevall A. Antibody-mediated immunity against tuberculosis: implications for vaccine development. Cell Host Microbe. 2013;13:250–262.
17. Verver S, Warren RM, Beyers N, et al. Rate of reinfection tuberculosis after successful treatment is higher than rate of new tuberculosis. Am J Respir Crit Care Med. 2005;171:1430–1435.
18. Marx FM, Dunbar R, Enarson DA, et al. The temporal dynamics of relapse and reinfection tuberculosis after successful treatment: a retrospective cohort study. Clin Infect Dis. 2014;58:1676–1683.
19. Chiang CY, Riley LW. Exogenous reinfection in tuberculosis. Lancet Infect Dis. 2005;5:629–636.
20. Glynn JR, Murray J, Bester A, Nelson G, Shearer S, Sonnenberg P. High rates of recurrence in HIV-infected and HIV-uninfected patients with tuberculosis. J Infect Dis. 2010;201:704–711.
21. Fox KK, Thomas JC, Weiner DH, Davis RH, Sparring PF, Cohen MS. Longitudinal evaluation of serovar-specific immunity to Neisseria gonorrhoeae. Am J Epidemiol. 1999;149:353–358.
22. Plummer FA, Simonsen JN, Chubb H, et al. Epidemiologic evidence for the development of serovar-specific immunity after gonococcal infection. J Clin Invest. 1989;83:1472–1476.
23. Rogawski McQuade ET, Liu J, Kang G, et al. Protection from natural immunity against enteric infections and etiology-specific diarrhea in a longitudinal birth cohort. J Infect Dis. 2020;222:1858–1868.
24. Tait DR, Hatherill M, Van Der Meeren O, et al. Final analysis of a trial of M72/AS01E vaccine to prevent tuberculosis. N Engl J Med. 2019:1–12.
25. Petousis-Harris H, Paynter J, Morgan J, et al. Effectiveness of a group B outer membrane vesicle meningococcal vaccine against gonorrhoea in New Zealand: a retrospective case-control study. Lancet. 2011;377:1603–1611.
26. Bourgeois AL, Wierzba TF, Walker RI. Status of vaccine research and development for enterotoxigenic Escherichia coli. Vaccine. 2016;34:2880–2886.
27. Dakin CJ, Numa AH, Wang H, Morton JR, Vertzayas CC, Henry RL. Inflammation, infection, and pulmonary function in infants and young children with cystic fibrosis. Am J Respir Crit Care Med. 2002;165:904–910.
28. Pillarissetti N, Williamson E, Linnane B, et al; Australian Respiratory Early Surveillance Team for Cystic Fibrosis (AREST CF). Infection, inflammation, and lung function decline in infants with cystic fibrosis. Am J Respir Crit Care Med. 2011;184:75–81.
29. Dagan R, Pelton S, Bakaletz L, Cohen R. Prevention of early episodes of otitis media by pneumococcal vaccines might reduce progression to complex disease. Lancet Infect Dis. 2016;16:480–492.
30. Howie VM, Plassouard JH, Sloyer J. The “otitis-prone” condition. Am J Dis Child. 1975;129:676–678.
31. Keusche GT, Denno DM, Black RE, et al. Environmental enteric dysfunction: pathogenesis, diagnosis, and clinical consequences. Clin Infect Dis. 2014;59(Suppl 4):S207–S212.
32. Mostofsky E, Coull BA, Mittleman MA. Analysis of observational self-matched data to examine acute triggers of outcome events with abrupt onset. Epidemiology. 2018;29:804–816.
33. Cochran WG. Some methods for strengthening the common χ² tests. Biometrics. 1954;10:417.
34. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959;22:719–748.
APPENDIX A: COMPETING RISKS

To derive equations (1)–(4), consider two competing, independent times \( \tau_i \) and \( \tau_j \) for events occurring at rates \( \lambda_i \) and \( \lambda_j \). We may obtain the probability for \( \tau_j \) to precede \( \tau_k \) from the improper integral

\[
Pr(\tau_j < \tau_k) = \int_0^\infty Pr(\tau_k > i) Pr(\tau_j = i) \, \text{d}i. \tag{12}
\]

Under the assumption of exponentially distributed event times,

\[
Pr(\tau_j < \tau_k) = \int_0^\infty \exp(-\lambda_j t) \lambda_j \exp(-\lambda_j t) \, \text{d}t = \frac{\lambda_j}{\lambda_j + \lambda_k}. \tag{13}
\]

APPENDIX B: TRUNCATION OF OBSERVATIONS

This appendix supports equations (6) and (7). The derivation above considers the improper integral

\[
Pr(\tau_j < \tau_k) = \int_0^\infty Pr(\tau_k > i) Pr(\tau_j = i) \, \text{d}i \tag{14}
\]

We obtain the same results when observations end at some time \( \delta \), more in line with the conduct of real-world studies:

\[
Pr(\tau_j < \tau_k) = \int_0^\delta Pr(\tau_k > i) Pr(\tau_j = i) \, \text{d}i = \frac{\delta \left(1 - \exp\left[-\delta(\lambda_j + \lambda_k)\right]\right)}{\lambda_j + \lambda_k}. \tag{15}
\]

Thus,

\[
Pr(E) = \frac{\lambda_p (1 - \exp[-\delta(\lambda_p + \lambda_N)])}{\lambda_p + \lambda_N} \tag{16}
\]

\[
Pr(F) = \frac{\lambda_N (1 - \exp[-\delta(\lambda_p + \lambda_N)])}{\lambda_p + \lambda_N} \tag{17}
\]

\[
Pr(G) = \frac{\theta \lambda_p (1 - \exp[-\delta(\theta \lambda_p + \lambda_N)])}{\theta \lambda_p + \lambda_N} \tag{18}
\]

\[
Pr(H) = \frac{\lambda_N (1 - \exp[-\delta(\lambda_N + \lambda_N)])}{\lambda_N + \lambda_N} \tag{19}
\]

with the additional terms canceling out in the matched odds ratio formulation:

\[
E(OR_{MH}) = \frac{\sum_i Pr(F_i) Pr(G_i)}{\sum_j Pr(E_j) Pr(H_j)} \tag{20}
\]

\[
= \sum \left(\frac{\lambda_p (1 - \exp[-\delta(\lambda_p + \lambda_N)])}{\lambda_p + \lambda_N} \times \frac{\theta \lambda_p (1 - \exp[-\delta(\theta \lambda_p + \lambda_N)])}{\theta \lambda_p + \lambda_N}\right) \tag{21}
\]

\[= \theta.
\]