Type: Original Research Article

Title: Inference of naturally-acquired immunity using a self-matched negative control design

Authors: Graham R. Northrup[1], Lei Qian[2], Katia Bruxvoort[2], Florian M. Marx[3,4], Lilith K. Whittles[5,6,7], Joseph A. Lewnard[1,8,9,*]

1. Center for Computational Biology, College of Engineering, University of California, Berkeley, Berkeley, California
2. Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, California
3. Desmond Tutu Tuberculosis Centre, Department of Paediatrics and Child Health, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
4. DST-NRF South African Centre of Excellence and Epidemiological Modelling and Analysis, Stellenbosch University, Stellenbosch, South Africa
5. Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, UK
6. MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, London, UK
7. NIHR Health Protection Research Unit in Modelling Methodology, School of Public Health, Imperial College London, London, UK
8. Division of Epidemiology, School of Public Health, University of California, Berkeley, Berkeley, California
9. Division of Infectious Diseases and Vaccinology, School of Public Health, University of California, Berkeley, Berkeley, California

* Corresponding author
Joseph A. Lewnard, PhD
Division of Epidemiology
School of Public Health
University of California, Berkeley
Room 5410
2121 Berkeley Way
Berkeley, California 94720
United States

Conflicts of interest: The authors declare no conflicts of interest.

Source of funding: This work was supported by the Dr. E. Dowdle fund from the University of California, Berkeley to JAL.

Materials availability: Code for replication of the analyses is available at github.com/gnorthrup/SelfMatchedNegativeControl.

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 re-estimate 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.

KEYWORDS

Natural immunity; frailty; negative control; hazard; self-matched; Mantel-Haenszel

INTRODUCTION

Host adaptive immune responses often protect against infection or disease when a pathogen is repeatedly encountered. Vaccines aim to exploit this mechanism of protection by exposing hosts to an attenuated infection, or to immunizing subunits of a pathogen. As such, evidence of protective naturally-acquired immunity provides 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 representing person-time at risk in the presence and absence of prior infection.[4–10] Thus, inference centers on the distribution of recurrent event times. Unmeasured heterogeneity in individuals' hazard rates of infection or disease presents a challenge in such analyses, originally termed a problem of "varying liabilities" by Greenwood and Yule[11] 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 analyses.[14] For instance, in studies of naturally-acquired immunity, 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 are thought to generate imperfect protection. Tuberculosis presents a notable example, where despite evidence of protective cell-mediated and humoral immunity,[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.[17–20] Similar conflict about the consequences of prior infection has arisen in epidemiologic studies of gonorrhea.[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 *Shigella* spp., *Campylobacter* spp., and various diarrheagenic *Escherichia coli* strains.[23] Evidence supporting the feasibility of protective vaccines against many of these pathogens suggests a need to revisit the impacts of naturally-acquired immunity.[24–26] Similar causal inference challenges arise in the relationship between chronic inflammation and repeated infection in conditions such as cystic fibrosis,[27,28] otitis media,[29,30] and environmental enteric dysfunction.[31]

Formalizing unmeasured heterogeneity as a problem of confounding suggests potential strategies to identify naturally-acquired protection. Terming Y_1 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 estimation of the causal relationship of interest, $Y_1 \rightarrow Y_2$. Conditioning on unmeasured individual factors by comparing observations during counterfactual risk periods from the same individual ($Y_1 \leftarrow \boxed{U} \rightarrow Y_2$) 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 in recent years.[32]

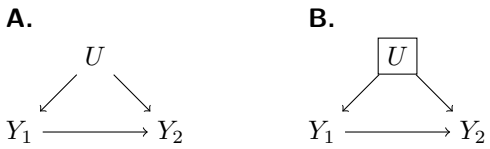

Figure 1: Directed acyclic graph addressing unmeasured confounding. We illustrate a causal framework wherein the effect of previous infection on time to subsequent infection ($Y_1 \rightarrow Y_2$) is of interest for analysis. One or more unmeasured confounding factors (U) creates a backdoor path (**A**) which can be blocked by conditioning on U (**B**).

In this paper, 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})[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.[4,5]

APPROACH

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 a competing risks estimation framework as we detail below. Last, let $X_i = 1$ indicate an individual has previously experienced infection with the pathogen of interest, and with $X_i = 0$ indicating the individual has no history of infection with the pathogen of interest.

Table 1: Parameters and definitions.

Parameter	Definition
λ_{Pi}	Rate at which individual *i* experiences a pre-specified clinical endpoint due to the pathogen of interest ("outcome of interest"), in the absence of naturally-acquired immunity
λ_{Ni}	Rate at which individual *i* experiences a negative-control outcome
θ	Hazard ratio for the outcome of interest, owing to naturally-acquired protection
$\beta_{Pi}(1)/\beta_{Pi}(0)$	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
$\beta_{Ni}(1)/\beta_{Ni}(0)$	Hazard ratio for the negative control outcome during the period after primary infection, relative to the period before primary infection, for individual *i*

Define A_i-D_i as random variables indicating event times for observations of $Y_i = 1$ and $Y_i = 0$, conditioned on X_i, according to the contingency structure presented in **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.

Table 2: Contingency table for event time distributions, given prior infection.

Exposure status		Outcome status	
		Outcome of interest $Y_i = 1$	Negative control outcome $Y_i = 0$
Previously uninfected	$X_i = 0$	$A_i \sim \mathrm{Exp}(\lambda_{Pi})$	$B_i \sim \mathrm{Exp}(\lambda_{Ni})$
Previously infected	$X_i = 1$	$C_i \sim \mathrm{Exp}(\theta\lambda_{Pi})$	$D_i \sim \mathrm{Exp}(\lambda_{Ni})$

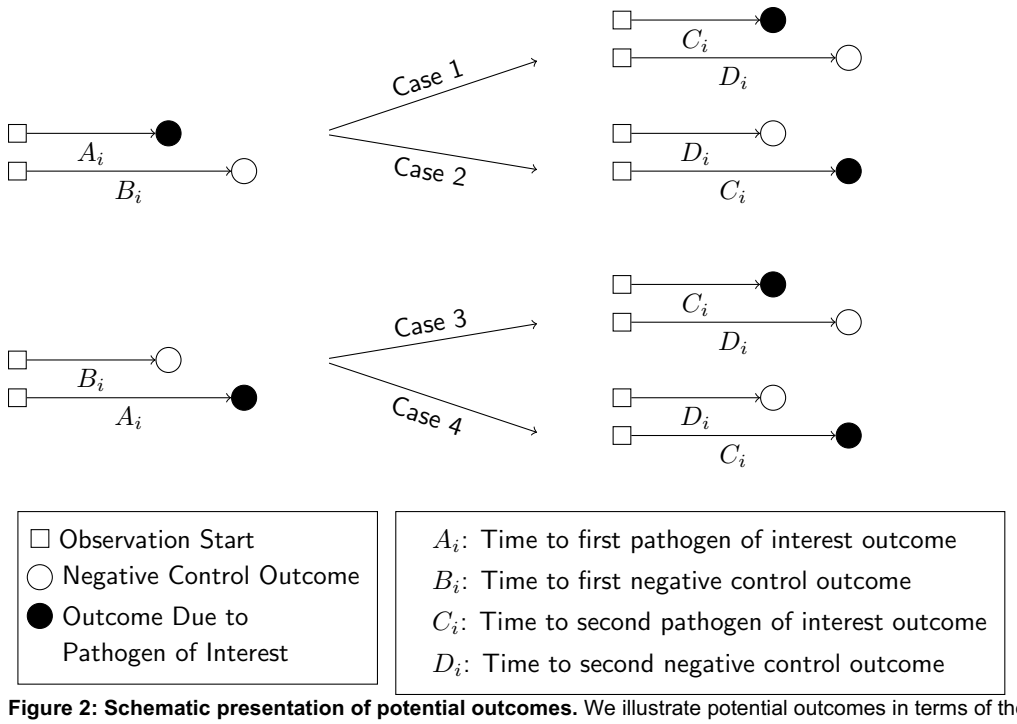

Figure 2: Schematic presentation of potential outcomes. We illustrate potential outcomes in terms of the sequence of events A_i-D_i for a given individual. In cases 1 and 2, we observe $A_i < B_i$ (truncating observation of a negative-control event while $X_i = 0$). In cases 3 and 4, we observe $B_i < A_i$, 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.

Event time distributions

Define the hazard rate at which an individual *i* experiences the outcome of interest as λ_{Pi}, and define θ as the hazard ratio of incidence of this outcome given previous infection (**Table 1**). Assuming events occur independently any time during the follow up, conditioning λ_{Pi} and θ, event times are exponentially distributed. Thus, the cumulative probability of experiencing the outcome by time *t*, for a previously-uninfected individual, is $1 - \exp(-\lambda_{Pi}t)$, while the cumulative probability of experiencing the outcome by time *t*, had the same individual counterfactually been previously infected, is $1 - \exp(-\theta\lambda_{Pi}t)$.

Consider that data are collected from each individual for endpoints besides the primary outcome of interest. Among these, suppose a negative control outcome occurs at a rate λ_{Ni} 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.[35] Under the same assumptions, the probability of experiencing the negative control outcome by time *t*, for individual *i*, is $1 - \exp(-\lambda_{Ni}t)$.

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 events E_i-H_i may be defined to indicate the relative ordering of event times A_i-D_i according to the contingency structure presented in **Table 3**. 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, where the negative-control outcome precedes infection or disease with the pathogen of interest while $X_i = 0$ and $X_i = 1$, respectively.

Table 3: Contingency table for competing risks, given prior infection.

Exposure status		**Outcome status**	
		Outcome of interest precedes negative control outcome $t\|Y_i = 1 < t\|Y_i = 0$	Negative control outcome precedes outcome of interest $t\|Y_i = 0 < t\|Y_i = 1$
Previously uninfected	$X_i = 0$	$E_i = A_i \leq B_i$	$F_i = B_i < A_i$
Previously infected	$X_i = 1$	$G_i = C_i \leq D_i$	$H_i = D_i < C_i$

Under the scenario of exponentially-distributed event times (formulated in **Appendices A and B**), we have

$$\Pr(E_i) = \Pr(A_i \leq B_i) = \frac{\lambda_{Pi}}{\lambda_{Pi} + \lambda_{Ni}} \tag{1}$$

$$\Pr(F_i) = \Pr(B_i < A_i) = \frac{\lambda_{Ni}}{\lambda_{Pi} + \lambda_{Ni}} \tag{2}$$

$$\Pr(G_i) = \Pr(C_i \leq D_i) = \frac{\theta\lambda_{Pi}}{\theta\lambda_{Pi} + \lambda_{Ni}} \tag{3}$$

$$\Pr(H_i) = \Pr(D_i < C_i) = \frac{\lambda_{Ni}}{\theta\lambda_{Pi} + \lambda_{Ni}}. \tag{4}$$

Consider the Mantel-Haenszel odds ratio[33,34] constructed from the competing risks of $Y_i = 1$ and $Y_i = 0$, given X_i, matching observations from each individual *i*:

$$OR_{MH} = \frac{\sum_i \mathrm{I}(F_i)\mathrm{I}(G_i)}{\sum_i \mathrm{I}(E_i)\,\mathrm{I}(H_i)}, \tag{5}$$

such that

$$\mathrm{E}(OR_{MH}) = \frac{\sum_i \Pr(F_i)\,\Pr\,(G_i)}{\sum_i \Pr(E_i)\,\Pr\,(H_i)}. \tag{6}$$

Using the above derivations of $\Pr(E_i)$ through $\Pr(H_i)$,

$$\mathrm{E}(OR_{MH}) = \frac{\sum_i \left(\frac{\lambda_{Ni}}{\lambda_{Pi} + \lambda_{Ni}} \times \frac{\theta\lambda_{Pi}}{\theta\lambda_{Pi} + \lambda_{Ni}}\right)}{\sum_i \left(\frac{\lambda_{Pi}}{\lambda_{Pi} + \lambda_{Ni}} \times \frac{\lambda_{Ni}}{\theta\lambda_{Pi} + \lambda_{Ni}}\right)} = \theta \frac{\sum_i \left(\frac{\lambda_{Ni}}{\lambda_{Pi} + \lambda_{Ni}} \times \frac{\lambda_{Pi}}{\theta\lambda_{Pi} + \lambda_{Ni}}\right)}{\sum_i \left(\frac{\lambda_{Pi}}{\lambda_{Pi} + \lambda_{Ni}} \times \frac{\lambda_{Ni}}{\theta\lambda_{Pi} + \lambda_{Ni}}\right)} = \theta. \tag{7}$$

Thus, the ratio of the matched odds for the outcome of interest to precede a negative-control outcome, given an individual's history of prior infection, provides an unbiased estimate of the effect of previous infection on time to recurrence of the outcome of interest.

Further Considerations

At a design level, self-matched inference reduces or eliminates the potential for bias due to time-invariant factors that individually influence risk.[36] However, complications arise when individuals' risk of experiencing these endpoints differs substantially during the periods before and after individuals experience their first infection with the pathogen of interest.

To consider the implications of such time-varying confounding, define $\beta_{Pi}(0)$ and $\beta_{Pi}(1)$ as multipliers on the rate of infection with the pathogen of interest when $X_i = 0$ and $X_i = 1$, respectively, due to all factors other than infection-derived immunity against the pathogen of interest (**Table 2**). Similarly, define $\beta_{Ni}(0)$and $\beta_{Ni}(1)$ as multipliers on the rate of the negative-control condition when $X_i = 0$ and $X_i = 1$, respectively, due to factors other than prior exposure to the pathogen of interest (**Table 4**). Here, $\theta\beta_{Pi}(1)/\beta_{Pi}(0)$ and $\beta_{Ni}(1)/\beta_{Ni}(0)$ are hazard ratios describing the relative incidence rate of the outcome of interest and the negative-control outcome, respectively, in the periods before and after infection with the pathogen of interest.

Table 4: Contingency table for event time distributions, given prior infection, under the scenario of time-variant confounding.

Exposure status		**Outcome status**	
		Outcome of interest $Y_i = 1$	Negative control outcome $Y_i = 0$
Previously uninfected	$X_i = 0$	$A_i \sim \text{Exp}(\beta_{Pi}(0)\lambda_{Pi})$	$B_i \sim \text{Exp}(\beta_{Ni}(0)\lambda_{Ni})$
Previously infected	$X_i = 1$	$C_i \sim \text{Exp}(\beta_{Pi}(1)\theta\lambda_{Pi})$	$D_i \sim \text{Exp}(\beta_{Ni}(1)\lambda_{Ni})$

Here the matched odds ratio is

$$\text{E}(OR_{MH}) = \frac{\sum_i \left(\frac{\beta_{Ni}(0)\lambda_{Ni}}{\beta_{Pi}(0)\lambda_{Pi} + \beta_{Ni}(0)\lambda_{Ni}} \times \frac{\theta\beta_{Pi}(1)\lambda_{Pi}}{\theta\beta_{Pi}(1)\lambda_{Pi} + \beta_{Pi}(1)\lambda_{Ni}} \right)}{\sum_i \left(\frac{\beta_{Pi}(0)\lambda_{Pi}}{\beta_{Pi}(0)\lambda_{Pi} + \beta_{Ni}(0)\lambda_{Ni}} \times \frac{\beta_{Ni}(1)\lambda_{Ni}}{\theta\beta_{Pi}(1)\lambda_{Pi} + \beta_{Pi}(1)\lambda_{Ni}} \right)}, \quad (8)$$

which reduces to θ only when $\beta_{Pi}(1)/\beta_{Pi}(0) = \beta_{Ni}(1)/\beta_{Ni}(0)$ for all individuals. As we address in the **Discussion**, this circumstance motivates the selection of negative-control outcomes which resemble the outcome of interest in their association with time-varying confounders such as individuals' age, health status, and sociodemographic exposures.

We may also consider a scenario where infection with the pathogen of interest alters individuals' risk of the negative control endpoint, in addition their risk of reinfection or recurrent disease due to the pathogen of interest. Such a circumstance may arise if infection with the pathogen of interest causes individuals to modify risk behaviors that affect multiple outcomes, or confers broad (e.g. multi-pathogen) immunity.[37,38]

Defining ρ as the outcome-agnostic effect of infection with the pathogen of interest on future outcomes yields the contingency structure of **Table 5**.

Table 5: Two-by-two table for event time distributions, given prior infection, in the presence of pathogen-agnostic effects of previous infection.

Exposure status		**Outcome status**	
		Outcome of interest $Y_i = 1$	Negative control outcome $Y_i = 0$
Previously uninfected	$X_i = 0$	$A_i \sim \text{Exp}(\lambda_{Pi})$	$B_i \sim \text{Exp}(\lambda_{Ni})$
Previously infected	$X_i = 1$	$C_i \sim \text{Exp}(\rho\theta\lambda_{Pi})$	$D_i \sim \text{Exp}(\rho\lambda_{Ni})$

Here,

$$\Pr(G_i) = \frac{\rho\theta\lambda_{Pi}}{\rho(\theta\lambda_{Pi} + \lambda_{Ni})} = \frac{\theta\lambda_{Pi}}{\theta\lambda_{Pi} + \lambda_{Ni}} \quad (9)$$

$$\Pr(H_i) = \frac{\rho\lambda_{Ni}}{\rho(\theta\lambda_{Pi} + \lambda_{Ni})} = \frac{\lambda_{Ni}}{\theta\lambda_{Pi} + \lambda_{Ni}} \quad (10)$$

so the parameterization of $\text{E}(OR_{MH}) = \frac{\sum_i \Pr(F_i)\Pr(G_i)}{\sum_i \Pr(E_i)\Pr(H_i)} = \theta$ is unchanged. This circumstance is analogous to the more general case where $\beta_{Pi}(1)/\beta_{Pi}(0) = \beta_{Ni}(1)/\beta_{Ni}(0) = \rho$.

COMPARISON TO COHORT DESIGN USING PROPORTIONAL HAZARDS ANALYSIS

Simulation study

We conducted a simulation study across various underlying distributions of λ_{Pi} and λ_{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 hazard rates among individuals with differing exposure or susceptibility to the outcome of interest:

1. "Naïve" proportional hazards model without inclusion of additional terms to account for differences in event times among individuals. We define the hazard ratio estimated via fitting this model as $\hat{\theta}_{\text{Naive}}$.

2. Proportional hazards model accounting for variation in individual frailty via "random effects". Fitting this model estimates the hazard ratio $\hat{\theta}_{\text{RE}}$ for the effect of previous infection, as well as $\hat{\sigma}^2$ representing the estimated variance in (log) individual-specific event rates, assumed to represent independent draws from a Normal distribution with mean 0.[39,40]

3. Proportional hazards model including Gamma-distributed frailty terms.[13] Fitting this model estimates the hazard ratio $\hat{\theta}_{\text{Frailty}}$ 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 estimates a hazard ratio $\hat{\theta}_{\text{FE}}$ for the effect of previous infection and estimates subject-specific rates of infection (via individual-specific intercepts) which have no pre-specified distributional assumption.

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

1. Truncated Normal distribution (with a pre-specified lower bound at $a = 0$);

2. Truncated Cauchy distribution (with a pre-specified 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 6**), 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 θ. We illustrate the distributions in **Figure 3**. Considering cohorts of 500 individuals, we drew λ_{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 λ_{Pi} values independently for each simulation. We used the simulated datasets to estimate $\hat{\theta}_{\text{Naive}}$, $\hat{\theta}_{\text{RE}}$, $\hat{\theta}_{\text{FE}}$, and $\hat{\theta}_{\text{Frailty}}$, taking the average of estimates obtained across all 500 iterations to obtain a single point estimate for each parameterization.

Table 6: Event rate distributions applied to simulation study.

Distribution	Parameters	Parameterizations[1]				
		I	II	III	IV	V
Truncated Normal	Mean μ	$\mu = 1$	$\mu = 1$	$\mu = 1$	$\mu = 1$	$\mu = 1$
	Variance σ^2	$\sigma = 1/4$	$\sigma = 1/2$	$\sigma = 1$	$\sigma = 2$	$\sigma = 4$
	Lower bound a	$a = 0$	$a = 0$	$a = 0$	$a = 0$	$a = 0$
	Upper bound b	$b = \infty$	$b = \infty$	$b = \infty$	$b = \infty$	$b = \infty$
Truncated Cauchy	Location x_0	$x_0 = 1$	$x_0 = 1$	$x_0 = 1$	$x_0 = 1$	$x_0 = 1$
	Scale γ	$\gamma = 1/8$	$\gamma = 1/4$	$\gamma = 1$	$\gamma = 4$	$\gamma = 8$
	Lower bound a	$a = 0$	$a = 0$	$a = 0$	$a = 0$	$a = 0$
	Upper bound b	$b = \infty$	$b = \infty$	$b = \infty$	$b = \infty$	$b = \infty$
Uniform	Lower bound a	$a = 7/8$	$a = 3/4$	$a = 1/2$	$a = 1/4$	$a = 0$
	Upper bound b	$b = 9/8$	$b = 5/4$	$b = 3/2$	$b = 7/4$	$b = 2$
Gamma	Shape k	$k = 8$	$k = 4$	$k = 1$	$k = 1/4$	$k = 1/8$
	Scale θ	$k = 1/8$	$k = 1/4$	$k = 1$	$k = 4$	$k = 8$
Gamma mixture (i)	Shapes k_1, k_2	$k_1 = 1/8$	$k_1 = 1/8$	$k_1 = 1/8$	$k_1 = 1/8$	$k_1 = 1/8$
		$k_2 = 3/8$	$k_2 = 15/8$	$k_2 = 15/16$	$k_2 = 3/32$	$k_2 = 3/320$
	Scale θ_1, θ_2	$\theta_1 = 1$	$\theta_1 = 1$	$\theta_1 = 1$	$\theta_1 = 1$	$\theta_1 = 1$
		$\theta_2 = 1/5$	$\theta_2 = 1$	$\theta_2 = 2$	$\theta_2 = 20$	$\theta_2 = 200$
	Weight[2] ω	$\omega = 0.5$	$\omega = 0.5$	$\omega = 0.5$	$\omega = 0.5$	$\omega = 0.5$
Gamma mixture (ii)	Shapes k_1, k_2	$k_1 = 1/2$	$k_1 = 1/2$	$k_1 = 1/2$	$k_1 = 1/2$	$k_1 = 1/2$
		$k_2 = 3/10$	$k_2 = 3/2$	$k_2 = 3$	$k_2 = 3/10$	$k_2 = 3/10$
	Scale θ_1, θ_2	$\theta_1 = 1$	$\theta_1 = 1$	$\theta_1 = 1$	$\theta_1 = 1$	$\theta_1 = 1$
		$\theta_2 = 1/5$	$\theta_2 = 1$	$\theta_2 = 2$	$\theta_2 = 20$	$\theta_2 = 200$
	Weight[2] ω	$\omega = 0.5$	$\omega = 0.5$	$\omega = 0.5$	$\omega = 0.5$	$\omega = 0.5$

1. Parameterizations are listed in order of increasing variance from I to V.
2. The weight parameter of the Gamma mixture distribution indicates the proportion of individuals whose rates are parameterized according to k_1, θ_1; the proportion with rates parameterized according to k_2, θ_2 is $1 - \omega$.

To compute $\hat{\theta}_{\mathrm{MH}}$, we drew rates ($\lambda_{Ni}$) and event times for negative control observations from each subject, assuming event times were exponentially-distributed with respect to the underlying rates. To standardize comparisons of $\hat{\theta}_{\mathrm{MH}}$ under differing distributions of λ_{Pi}, we defined $\lambda_{Ni} = 1$ for all i under each simulation.

To investigate how the different modeling frameworks performed in capturing the distribution of individual-specific hazard rates, we saved estimates of individual-specific fixed effects, random effects, and frailties alongside estimates of $\hat{\theta}$. We fitted a single density kernel to the distribution of individual-specific estimates across 10 simulated cohorts for each true value of θ and underlying distribution of λ_{Pi}.

Results

We plot distributions and estimates under each approach in **Figure 3**. The naïve hazards ratio tended to overestimate θ, leading to under-estimation of the degree of protection ($1 - \theta$). Bias was minimized as θ approached zero, consistent with a scenario of strong protective immunity. Values of $\hat{\theta}_{\mathrm{Naive}}$ 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 due to the pathogen of interest, when in fact prior infection is protective. For all distributions considered, bias in $\hat{\theta}_{\mathrm{Naive}}$ was greatest under parameterizations yielding the highest between-individual variance in λ_{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 θ than $\hat{\theta}_{\mathrm{Naive}}$. 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 λ_{Pi} values were drawn from Gamma or Gamma mixture distributions, and tended to increase under distributions with greater variance in λ_{Pi}, or greater irregularity in the case of Gamma mixture distributions. In contrast, fixed-effects models estimating multipliers on hazard rates for each individual tended to under-estimate θ under most distributions of λ_{Pi}, although both $\hat{\theta}_{\mathrm{FE}} > \theta$ and $\hat{\theta}_{\mathrm{FE}} < \theta$ were apparent in simulations using the truncated Cauchy distribution for λ_{Pi}. For the truncated Normal distribution, bias in $\hat{\theta}_{\mathrm{FE}}$ decreased with greater

variance in λ_{Pi}, whereas for the Uniform, Gamma, and Gamma mixture distributions, bias increased with greater variance in λ_{Pi}.

Biased estimation of θ 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 θ and distributions of λ_{Pi} (**Supplemental Digital Content, Figures S1-S3**).

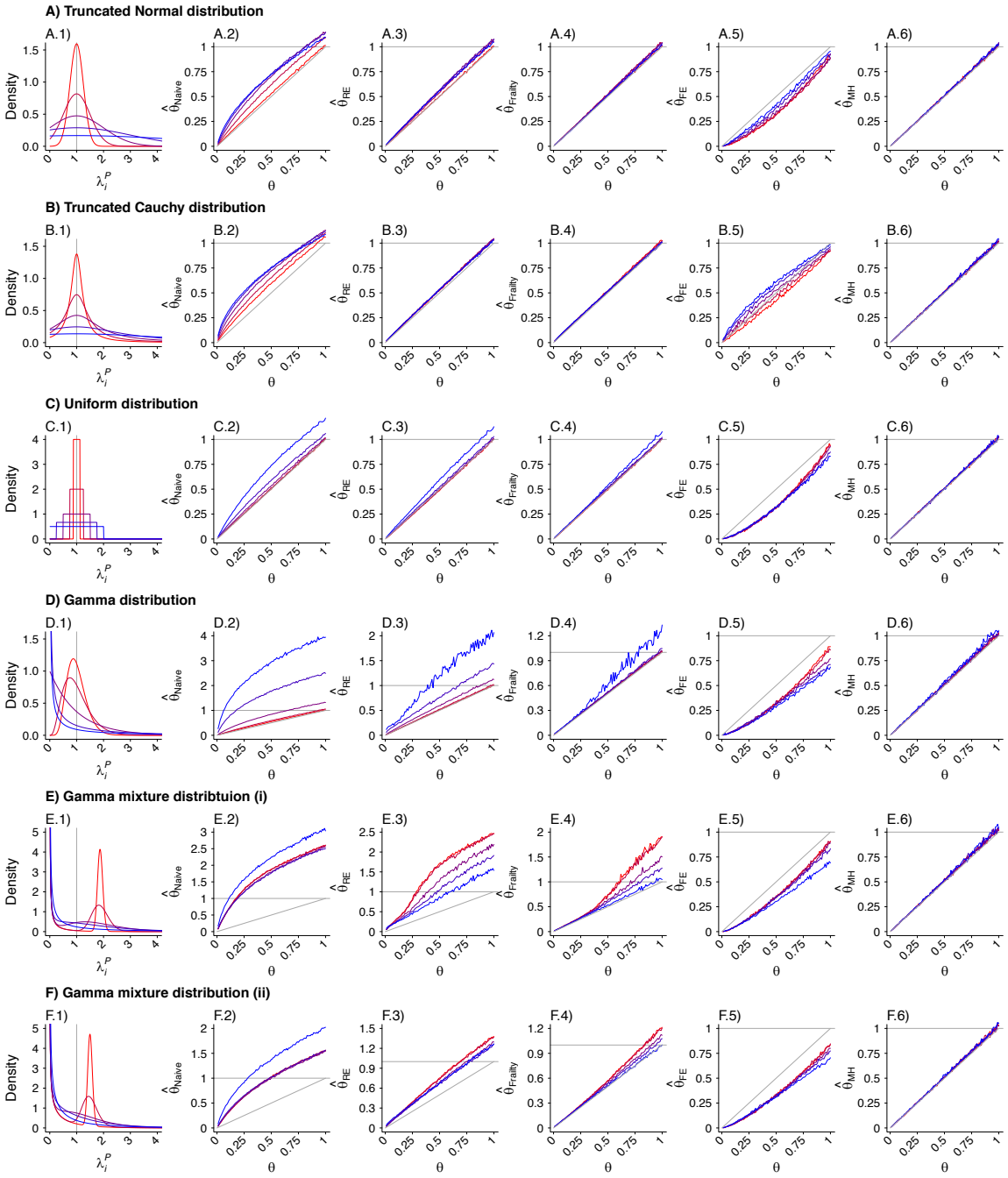

Figure 3: Simulated distributions and hazard ratio estimates under "naïve" inference approaches and under the proposed approach of self-matched inference with negative controls. Panels are organized to present, in each row, the assumed distribution (column 1), the estimate $\hat{\theta}_{\text{Naive}}$ 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 grey in columns 2-6 indicate where estimates would

recover the true value, i.e. $\hat{\theta} = \theta$. Horizontal grey 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.

SAMPLE SIZE CONSIDERATIONS

Simulation study

To inform applications of the proposed method, we next assessed statistical power under differing conditions. A test statistic (ξ_{MH}) has previously been identified for OR_{MH} under the null hypothesis of no difference in risk given exposure.[41] For the contingency structure (**Table 3**) formulated from the terms E_i-H_i (by which we define OR_{MH}), this statistic can be written generally as

$$\xi_{MH} = \frac{(\sum_i^N (E_i - \frac{(E_i + F_i)(E_i + G_i)}{2}))^2}{\sum_i^N \frac{(E_i + F_i)(E_i + G_i)(F_i + H_i)(G_i + H_I)}{4}}, \quad (11)$$

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,

$$\xi_{MH} = \frac{(\sum_i^N (E_i - G_i))^2}{\sum_i^N (E_i + G_i)(2 - E_i - G_i)}, \quad (12)$$

which is expected to follow a χ^2 distribution with one degree of freedom under the null hypothesis. We calculated values of ξ_{MH} obtained for cohorts of varying sizes under differing parameterizations of θ, λ_{Pi}, and λ_{Ni}. For values of $\theta \in \{0.1, 0.2, \ldots, 0.9\}$, we sampled individual event times A_i-D_i for a population of 100,000 individuals whom we subsequently partitioned (without replacement) into 2000 hypothetical study cohorts each of size N=25, 50, 75, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, and 1000. For these analyses, we considered λ_{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 6**. 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, i.e. $\Pr(\hat{\xi}_{MH} > \xi_{MH}^{97.5\%} = 5.02)$.

To assess how correlation between λ_{Pi} and λ_{Ni} could affect the statistical power of estimates, we conducted simulations under two sets of assumptions. Under the first, we considered $\lambda_{Ni} = 1$ for all i (equal to the expected value of λ_{Pi} under all parameterizations), so that $\lambda_{Ni} \perp \lambda_{Pi}$; 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 bound power estimates, corresponding to assuming no correlation and perfect correlation between λ_{Ni} and λ_{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 1000 subjects; statistical power for estimation of $\theta = 0.9$ was 10% or lower under nearly all conditions explored.

For simulations with $\lambda_{Ni} \perp \lambda_{Pi}$, statistical power was weaker under parameterizations resulting in greater variance in λ_{Pi}. In contrast, for simulations with $\lambda_{Ni} = \lambda_{Pi}$, differences in statistical power were less

strongly apparent with increasing variance in λ_{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.

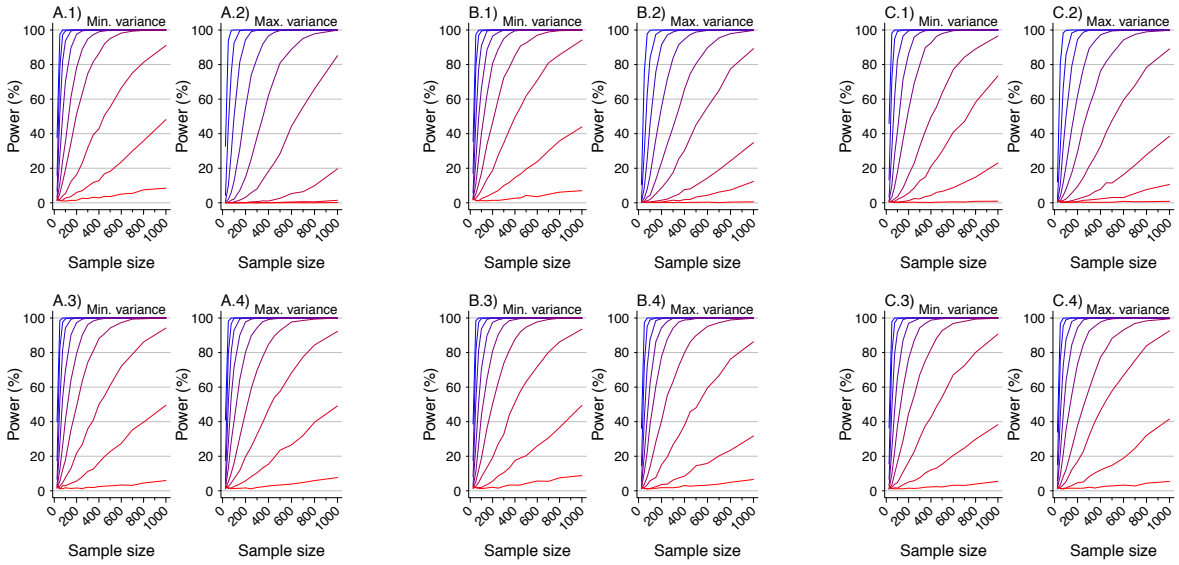

Figure 4: Statistical power for simulated analyses using the proposed approach of self-matched inference via negative controls. Each panel presents the statistical power for rejecting the null hypothesis with two-sided p<0.05 under varying conditions. Lines plotted in red to blue correspond to decreasing values of θ: 0.9 (red), 0.8, 0.7, …, 0.1 (blue), corresponding to increasing protection from 10% to 90%. Plots are presented in groups of 4 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 ($\lambda_i^N = 1 \forall i$). 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 is maximized ($\lambda_i^N = \lambda_i^P$). 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 (λ_i^P; 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 (λ_i^P; i.e., parameterization V in **Table 6**).

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 children in Vellore, India. These datasets have been described extensively in primary study publications[4,5] and subsequent re-analyses.[15,42] Similar designs were employed for the studies. Briefly, pregnant mothers were enrolled prior to childbirth, and children were followed from birth to ages 2 years (in Mexico City) and 3 years (in Vellore). Investigators aimed to identify all rotavirus infections through routine testing of asymptomatic stool specimens (collected by field workers at regular home visits) for rotavirus, and by monitoring children for anti-rotavirus seroconversion over serial blood draws at scheduled intervals. Active surveillance was undertaken for all cases of gastroenteritis among children to characterize symptoms and test diarrheal stool specimens for rotavirus.

Initial analyses of the datasets led to differing conclusions about the strength of protection against rotavirus gastroenteritis (RVGE). 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.[42] 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 re-estimate naturally-acquired protection against RVGE in the cohort datasets. Here, RVGE episodes (acute, new-onset diarrhea with rotavirus detected in the stool) are the outcome of interest and acute, new-onset diarrhea episodes without rotavirus detection as the negative control. 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 birth cohort studies.[4,5]

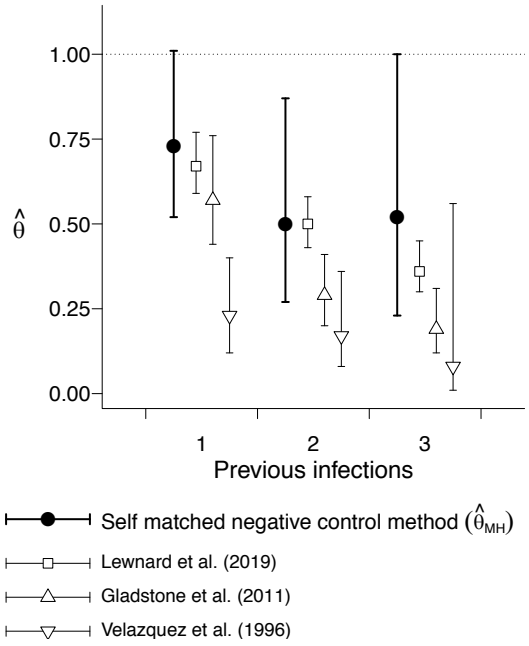

Figure 5: Estimated protection against rotavirus gastroenteritis associated with previous infection. We plot point estimates and 95% confidence intervals (lines) for estimates of the hazard ratio of rotavirus gastroenteritis associated with having previously experienced one, two, and three previous infections, versus zero previous infections, estimated via re-analysis of the Mexico City and Vellore rotavirus birth cohort studies.[4,5] Analyses include rotavirus-negative diarrhea occurrences as a negative control endpoint.

DISCUSSION

We propose a novel self-matched negative control method for estimating the hazard ratio of time to infection or disease due to a pathogen of interest, given previous infection. Analytically and via simulation, we show this method recovers unbiased estimates under a range of conditions, including when individual incidence rates of the outcome of interest are drawn from highly irregular or skewed distributions. We find these irregular or skewed distributions may lead to bias under proportional hazards models with commonly-used frailty estimation frameworks. Desirably, the proposed approach requires no parametric assumptions other than event-times being exponentially distributed with respect to their underlying,

individual-specific rates of occurrence. 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 negative-control outcome, λ_{Pi} and λ_{Ni}. "Test-negative" control conditions which resemble the outcome of interest, but are not attributable to the same pathogen,[43,44] may provide a compelling choice, particularly if their occurrence is predicted by similar risk factors. For instance, shared risk factors are well-documented for rotavirus-positive and rotavirus-negative diarrhea.[15,31] Considering respiratory illness, multiple etiologic viruses may share similar seasonal transmission patterns,[45] routes of transmission via high-risk contact,[46] and associations of disease progression or severity with host comorbidities.[47] For sexually-transmitted infections, particular risk behaviors[48] differing among individuals or over time could alter risk of any infection, rather than infection with the pathogen of interest alone.[25] In the context of real-world cohort studies, test-negative control conditions which 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.

APPENDIX A: COMPETING RISKS

For two competing, independent event times τ_j and τ_k occurring at rates r_j and r_k, the probability for τ_j to precede τ_k is

$$\Pr(\tau_j < \tau_k) = \int_0^\infty \Pr(\tau_k > t) \Pr(\tau_j = t)\, \mathrm{d}t. \tag{13}$$

Under the assumption of exponentially-distributed event times,

$$\Pr(\tau_j < \tau_k) = \int_0^\infty \exp(-r_k t)\, \mathrm{r_j} \exp(-r_j t)\, \mathrm{d}t = \frac{r_j}{r_j + r_k}. \tag{14}$$

APPENDIX B: TRUNCATION OF OBSERVATIONS

The derivation above considers the indefinite integral

$$\Pr(\tau_j < \tau_k) = \int_0^\infty \Pr(\tau_k > t) \Pr(\tau_j = t)\, \mathrm{d}t \tag{15}$$

We obtain the same results when considering bounded observations truncated at time δ, more in line with the conduct of real-world studies:

$$\Pr(\tau_j < \tau_k) = \int_0^\delta \Pr(\tau_k > t) \Pr(\tau_j = t)\, \mathrm{d}t = \frac{r_j\left(1 - \exp\left[-\delta(r_k + r_j)\right]\right)}{r_k + r_j}. \tag{16}$$

Thus,

$$\Pr(E_i) = \frac{\lambda_{Pi}(1 - \exp[-\delta(\lambda_{Pi} + \lambda_{Ni})])}{\lambda_{Pi} + \lambda_{Ni}} \tag{17}$$

$$\Pr(F_i) = \frac{\lambda_{Ni}(1 - \exp[-\delta(\lambda_{Pi} + \lambda_{Ni})])}{\lambda_{Pi} + \lambda_{Ni}} \tag{18}$$

$$\Pr(G_i) = \frac{\theta\lambda_{Pi}(1 - \exp[-\delta(\theta\lambda_{Pi} + \lambda_{Ni})])}{\theta\lambda_{Pi} + \lambda_{Ni}} \tag{19}$$

$$\Pr(H_i) = \frac{\lambda_{Ni}(1 - \exp[-\delta(\theta\lambda_{Pi} + \lambda_{Ni})])}{\theta\lambda_{Pi} + \lambda_{Ni}}, \tag{20}$$

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

$$\begin{aligned} \mathrm{E}(OR_{MH}) &= \frac{\sum_i \Pr(F_i)\Pr(G_i)}{\sum_i \Pr(E_i)\Pr(H_i)} \\ &= \frac{\sum_i \left(\frac{\lambda_{Ni}(1 - \exp[-\delta(\lambda_{Pi} + \lambda_{Ni})])}{\lambda_{Pi} + \lambda_{Ni}} \times \frac{\theta\lambda_{Pi}(1 - \exp[-\delta(\theta\lambda_{Pi} + \lambda_{Ni})])}{\theta\lambda_{Pi} + \lambda_{Ni}} \right)}{\sum_i \left(\frac{\lambda_{Pi}(1 - \exp[-\delta(\lambda_{Pi} + \lambda_{Ni})])}{\lambda_{Pi} + \lambda_{Ni}} \times \frac{\lambda_{Ni}(1 - \exp[-\delta(\theta\lambda_{Pi} + \lambda_{Ni})])}{\theta\lambda_{Pi} + \lambda_{Ni}} \right)} \\ &= \theta. \end{aligned} \tag{21}$$

REFERENCES

1. Lopman B, Kang G. In praise of birth cohorts: Norovirus infection, disease, and immunity. *Clin Infect Dis*. 2014;58(4):492-494.

2. Pasetti MF, Levine MM. Insights from natural infection-derived immunity to cholera instruct vaccine efforts. *Clin Vaccine Immunol*. 2012. doi:10.1128/CVI.00543-12

3. Heesterbeek H, Anderson RM, Andreasen V, et al. Modeling infectious disease dynamics in the complex landscape of global health. *Science*. 2015. doi:10.1126/science.aaa4339

4. Velázquez FR, Matson DO, Calva JJ, et al. Rotavirus infections in infants as protection against subsequent infections. *N Engl J Med*. 1996;335(14):1022-1028.

5. Gladstone BP, Ramani S, Mukhopadhya I, et al. Protective effect of natural rotavirus infection in an Indian birth cohort. *N Engl J Med*. 2011;365(4):337-346.

6. Rouhani S, Peñataro Yori P, Paredes Olortegui M, et al. Norovirus infection and acquired immunity in 8 countries: Results from the MAL-ED study. *Clin Infect Dis*. 2016;62(10):1210-1217.

7. King AA, Ionides EL, Pascual M, Bouma MJ. Inapparent infections and cholera dynamics. *Nature.* 2008;454:877-880.

8. Andrews JR, Noubary F, Walensky RP, Cerda R, Losina E, Horsburgh CR. Risk of progression to active tuberculosis following reinfection with *Mycobacterium tuberculosis*. *Clin Infect Dis*. 2012. doi:10.1093/cid/cir951

9. Katzelnick LC, Gresh L, Halloran ME, et al. Antibody-dependent enhancement of severe dengue disease in humans. *Science*. 2017. doi:10.1126/science.aan6836

10. Rodriguez-Barraquer I, Arinaitwe E, Jagannathan P, et al. Quantifying heterogeneous malaria exposure and clinical protection in a cohort of Ugandan children. *J Infect Dis.* 2016. doi:10.1093/infdis/jiw301

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(2):255.

12. Arbous AG, Kerrich JE. Accident statistics and the concept of accident-proneness. *Biometrics*. 1951. doi:10.2307/3001656

13. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*. 1979. doi:10.2307/2061224

14. Glynn RJ, Buring JE. Ways of measuring rates of recurrent events. *BMJ.* 1996. doi:10.1136/bmj.312.7027.364

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. doi:10.1371/journal.pcbi.1007014

16. Achkar JM, Casadevall A. Antibody-mediated immunity against tuberculosis: Implications for vaccine development. *Cell Host Microbe*. 2013. doi:10.1016/j.chom.2013.02.009

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. doi:10.1164/rccm.200409-1200OC

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(12):1676-1683.

19. Chiang CY, Riley LW. Exogenous reinfection in tuberculosis. *Lancet Infect Dis*. 2005;5(10):629-636. doi:10.1016/S1473-3099(05)70240-1

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. doi:10.1086/650529

21. Fox KK, Thomas JC, Weiner DH, Davis RH, Frederick Sparling P. Longitudinal evaluation of serovar-specific immunity to Neisseria gonorrhoeae. *Am J Epidemiol*. 1999. doi:10.1093/oxfordjournals.aje.a009820

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. doi:10.1172/JCI114040

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. doi:10.1093/infdis/jiaa031

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*. 2017;390(10102):1603-1610.

26. Mani S, Wierzba T, Walker RI. Status of vaccine research and development for *Shigella*. *Vaccine*. 2016. doi:10.1016/j.vaccine.2016.02.075

27. Dakin CJ, Numa AH, Wang H, Morton JR, Vertzyas CC, Henry RL. Inflammation, infection, and pulmonary function in infants and young children with cystic fibrosis. *Am J Respir Crit Care Med*. 2002. doi:10.1164/ajrccm.165.7.2010139

28. Pillarisetti N, Williamson E, Linnane B, et al. Infection, inflammation,and lung function decline in infants with cystic fibrosis. *Am J Respir Crit Care Med*. 2011. doi:10.1164/rccm.201011-1892OC

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(4):480-492.

30. Howie VM, Ploussard JH, Sloyer J. The “otitis-prone” condition. *Am J Dis Child*. 1975;129(6):676-678.

31. Keusch GT, Denno DM, Black RE, et al. Environmental enteric dysfunction: Pathogenesis, diagnosis, and clinical consequences. *Clin Infect Dis.* 2014;59: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(6):804-816.

33. Cochran WG. Some methods for strengthening the common χ^2 Tests. *Biometrics*. 1954;10(4):417.

34. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. 1959;22(4):719-748.

35. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative Controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*. 2010. doi:10.1097/EDE.0b013e3181d61eeb

36. Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: The self-controlled case series method. *Stat Med*. 2006. doi:10.1002/sim.2302

37. Cowling BJ, Nishiura H. Virus interference and estimates of influenza vaccine effectiveness from test-negative studies. *Epidemiology*. 2012. doi:10.1097/EDE.0b013e31826b300e

38. Gordon A, Gresh L, Ojeda S, et al. Prior dengue virus infection and risk of Zika: A pediatric cohort in Nicaragua. *PLoS Med*. 2019. doi:10.1371/journal.pmed.1002726

39. Pankratz VS, De Andrade M, Therneau TM. Random-effects cox proportional hazards model: General variance components methods for time-to-event data. *Genet Epidemiol*. 2005. doi:10.1002/gepi.20043

40. Therneau TM. Package “survival.” 2019. https://github.com/therneau/survival.

41. Mantel N. Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *J Am Stat Assoc*. 1963. doi:10.1080/01621459.1963.10500879

42. Lewnard JA, Lopman BA, Parashar UD, et al. Naturally acquired immunity against rotavirus infection and gastroenteritis in children: Paired reanalyses of birth cohort studies. *J Infect Dis*. 2017;216(3).

43. Lewnard JA, Tedijanto C, Cowling BJ, Lipsitch M. Measurement of vaccine direct effects under the test-negative design. *Am J Epidemiol*. 2018. doi:10.1093/aje/kwy163

44. Sullivan SG, Tchetgen EJT, Cowling BJ. Theoretical basis of the test-negative study design for assessment of influenza vaccine effectiveness. *Am J Epidemiol*. 2016;184(5):345-353.

45. Monto AS, Sullivan KM. Acute respiratory illness in the community: frequency of illness and the agents involved. *Epidemiol Infect*. 1993. doi:10.1017/S0950268800050779

46. Mossong J, Hens N, Jit M, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med*. 2008;5(3):0381-0391.

47. Fitzner J, Qasmieh S, Mounts AW, et al. Revision of clinical case definitions: Influenza-like illness and severe acute respiratory infection. *Bull World Health Organ*. 2018. doi:10.2471/BLT.17.194514

48. Gallo MF, Steiner MJ, Warner L, et al. Self-reported condom use is associated with reduced risk of chlamydia, gonorrhea, and trichomoniasis. *Sex Transm Dis*. 2007. doi:10.1097/OLQ.0b013e318073bd71

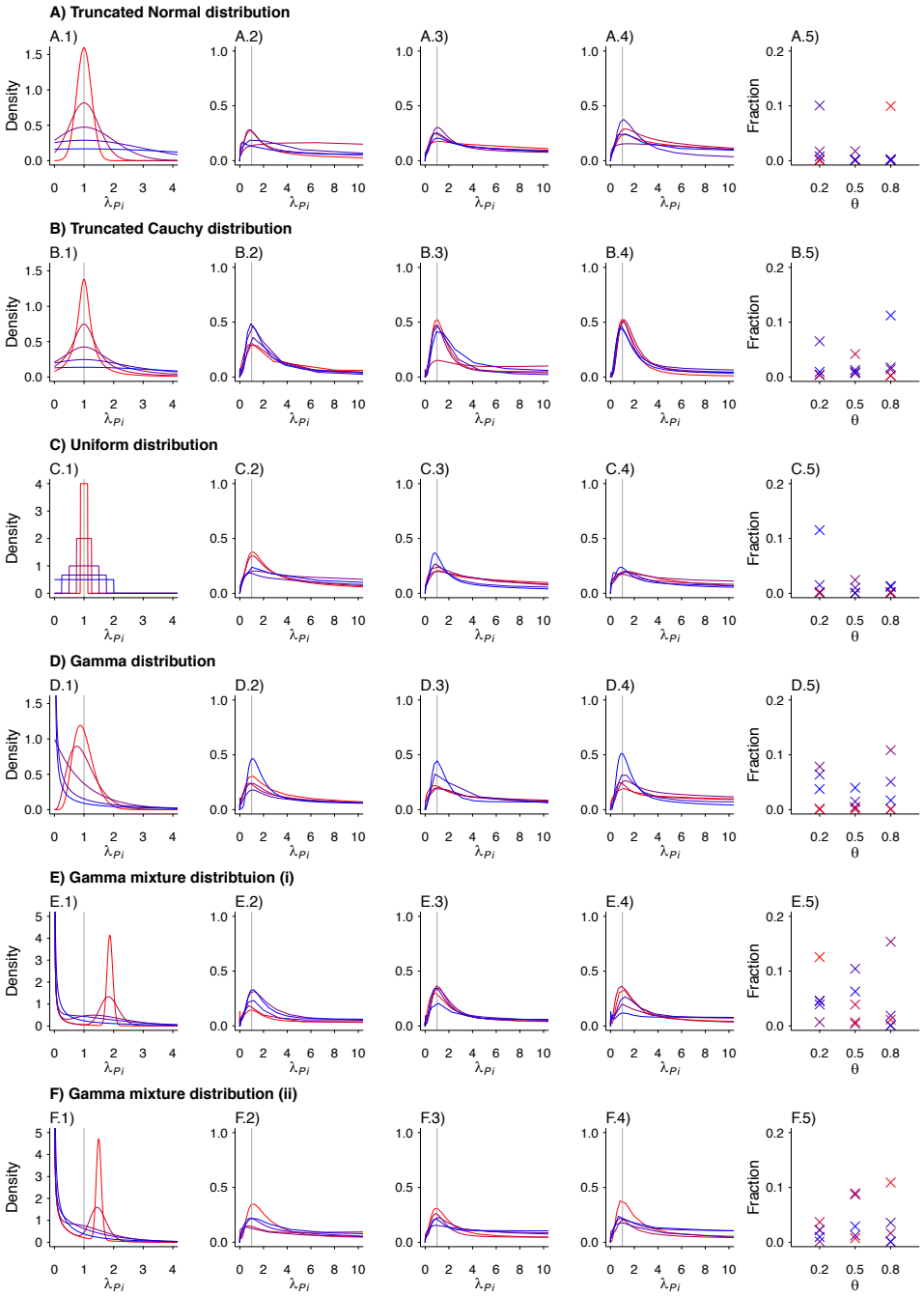

FIGURE S1: Estimated Density Kernels For Fixed Effects Model. We plot the estimated density kernels for the individual effects estimated by the fixed effects model for each individual in a simulated study. In each row, column 1 is a reproduction of the densities used to produce the individual λ_{Pi}. Columns 2-4 are the estimated density kernels for $\theta = 0.2, \theta = 0.5, \theta = 0.8$, respectively. Column 5 shows the proportion of estimates which were over 1000 for each set of parameters. 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.

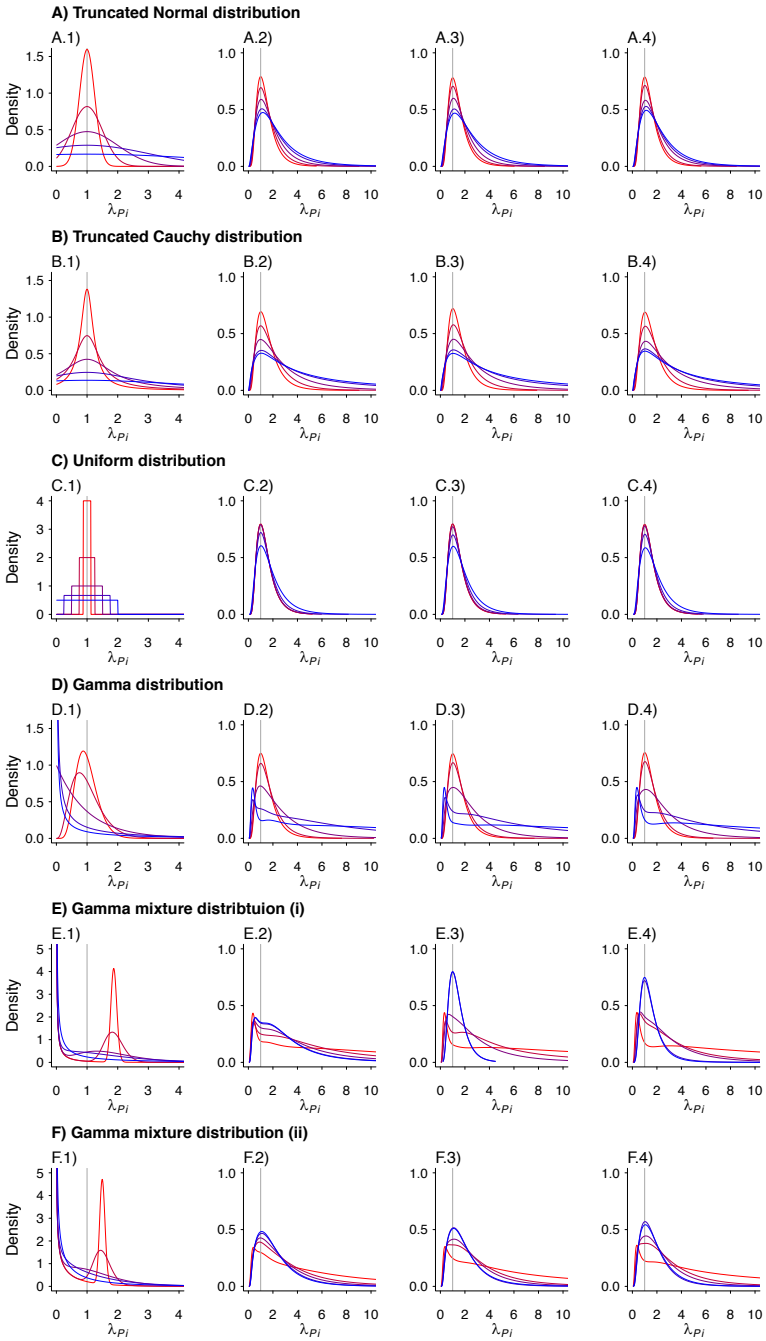

FIGURE S2: Estimated Density Kernels For Random Effects Model. We plot the estimated density kernels for the individual effects estimated by the random effects model for each individual in a simulated study. In each row, column 1 is a reproduction of the densities used to produce the individual λ_{Pi}. Columns 2-4 are the estimated density kernels for $\theta = 0.2$, $\theta = 0.5$, $\theta = 0.8$, respectively. 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.

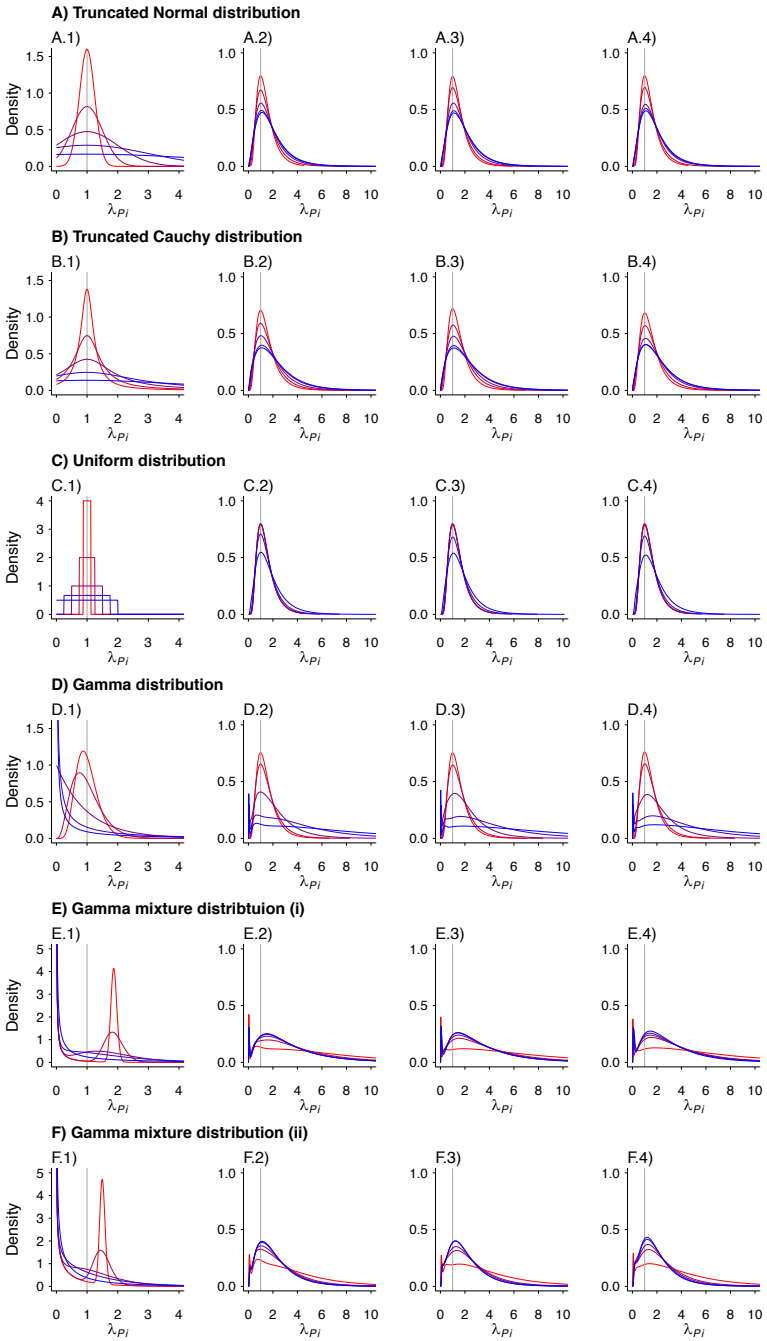

FIGURE S3: Estimated Density Kernels For Gamma Frailty Model. We plot the estimated density kernels for the individual effects estimated by the gamma frailty model for each individual in a simulated study. In each row, column 1 is a reproduction of the densities used to produce the individual λ_{Pi}. Columns 2-4 are the estimated density kernels for $\theta = 0.2$, $\theta = 0.5$, $\theta = 0.8$, respectively. 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.