Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

Estimating Risk Factors for Pathogenic Dose Accrual From Longitudinal Data

Daniel K. Sewell[1] | Kelly K. Baker[2,3]

[1]Department of Biostatistics, University of Iowa, Iowa City, Iowa, USA | [2]Department of Occupational and Environmental Health, University of Iowa, Iowa City, Iowa, USA | [3]Department of Epidemiology and Environmental Health, State University of New York at Buffalo, New York, New York, USA

Correspondence: Daniel K. Sewell (daniel-sewell@uiowa.edu)

Received: 20 December 2024 | **Revised:** 15 August 2025 | **Accepted:** 22 September 2025

Funding: This work was supported by Fogarty International Center, R01 TW011795.

Keywords: dose–response | incidence | infectious disease | quantitative microbial risk assessment

ABSTRACT

Estimating risk factors for the incidence of a disease is crucial for understanding its etiology. For diseases caused by enteric pathogens, off-the-shelf statistical model-based approaches do not consider the biological mechanisms through which infection occurs and thus can only be used to make comparatively weak statements about the association between risk factors and incidence. Building off of established work in quantitative microbiological risk assessment, we propose a new approach to determining the association between risk factors and dose accrual rates. Our more mechanistic approach achieves a higher degree of biological plausibility, incorporates currently ignored sources of variability, and provides regression parameters that are easily interpretable as the dose accrual rate ratio due to changes in the risk factors under study. We also describe a method for leveraging information across multiple pathogens. The proposed methods are available as an R package at https://github.com/dksewell/dare. Our simulation study shows unacceptable coverage rates from generalized linear models, while the proposed approach empirically maintains the nominal rate even when the model is misspecified. Finally, we demonstrated our proposed approach by applying our method to infant data obtained through the PATHOME study (https://reporter.nih.gov/project-details/10227256), discovering the impact of various environmental factors on infant enteric infections.

1 | Introduction

Enteric infections are a significant source of morbidity and mortality globally. The World Health Organization estimates that each year there are nearly 1.7 billion cases of childhood diarrheal disease, with over 440 000 deaths in children under 5 years old [1]. While this disproportionately impacts low- to middle-income countries, high-income countries, too, are highly impacted. For example, *Clostridioides difficile* infections alone affect roughly 500 000 individuals in the United States each year, leading to around 30 000 deaths [2]. Obtaining a deeper understanding of transmission dynamics and how they vary according to individual-level characteristics is critical to understanding disease etiology [3], which in turn leads to more informed intervention design and health policies.

1.1 | Incidence Estimation

Incidence is one of the most core quantitative epidemiological measures available to help understand infectious disease transmission dynamics [4]. The simplest approach to describing infection rates is through the incidence rate, defined to be the number of new cases over a specified time interval, or, relatedly,

through the incidence proportion given by the proportion of at-risk individuals to become infected over a specified time interval. Ascertaining how incidence varies in subpopulations associated with specific risk factors in this way is predicated on several factors, including a clear delineation of the population into subpopulations and population-level surveillance. Although the latter issue can be addressed through sampling techniques and established statistical inference procedures, the former leads to intractability for multiple risk factors, as this number grows exponentially with the number of individual-level factors. These two problems—the requirement to sample and having a large number of partitions of the population—synergize in that within each subpopulation, sufficient numbers of subjects must be recruited so as to reliably estimate the rate at which new cases occur. This is, of course, yet further exacerbated by rare diseases. To ameliorate these issues, generalized linear models (GLMs) based on the Poisson distribution have been used, which makes the additional assumption that the variation in incidence across subpopulations can (on the log scale) be described mathematically as an additive combination of the subpopulations' risk factors [5].

An additional problem with the above approaches arises when one or more individual-level risk factors are not categorical or finite in nature. In such a case, either arbitrary thresholding must occur, or some other model-based approach is taken. Several such approaches can be found in the extant literature. Some researchers turn to logistic regression (e.g., [6]), estimating the effect of individual-level risk factors on the change in odds of becoming newly infected over a specified time period. However, this approach requires that each individual recruited to the study be observed for the same amount of time, for example, 2 weeks. In some cases, this is feasible, but this is often not the case due to a variety of reasons (e.g., if study participants schedule their follow-up time within a varying time window from baseline, or if it is not possible to precisely schedule biological specimen collection). Another approach is the use of survival models, such as the Cox Proportional Hazards model (e.g., [7]). This approach, however, typically requires that the new infection causes acute symptoms which allow the infection to be surveilled passively, and that the incubation period is either known or negligible. Another approach which is sufficiently flexible to handle varying observation lengths is the use of the complementary log–log link function in a GLM based on the Bernoulli distribution (e.g., [8]). By using the log of the individuals' time intervals as an offset in the model, one can estimate the rate at which new cases develop for a given set of covariates.

1.2 | A Mechanistic View

There are several key steps by which microorganisms in the environment result in infecting an individual, each of which is an important source of variability. First, environmental and behavioral risk factors impact what an individual is exposed to. As a running example, consider access or lack of access a child has to a private latrine. Second, these risk factors lead to varying dose concentrations of the various fomites, vectors, and vehicles to which an individual is exposed, and combined with stochastic behaviors, such as the quantity of media consumed or the number of hand-to-mouth contacts, result in highly variable expected ingested doses. For example, even after conditioning on latrine access, the dose concentration of an enteric pathogen on a child's hands will vary stochastically based on who else has used the latrine recently, and what the child happens to touch and how often on a given day. A third source of variability is in the actual dose received given the expected ingested dose. For example, conditioning on the dose concentration on a child's hands and a certain number of hand-to-mouth contacts, the actual dose ingested will still vary due to heterogeneity of pathogen concentration on the hands' surfaces, which part of the hand is in contact with which part of the mouth and for how long. Additionally, temporal pathogen-environment interactions can enhance or reduce pathogen survival in the environment, affecting the viability of pathogen dose. Finally, pathogens have varying abilities to propagate within a host [9]. Infection "can be described by competing processes of birth and death within the host, infection resulting when birth is sufficient to produce a body burden above some critical level to induce the effect" ([10], p. 268). That is, conditioning on the dose ingested, each pathogen has a random chance at surviving to be able to cause an infection, affected by the host's capacity to mount a rapid and effective innate and adaptive immune response, including targeted antibodies developed from prior exposure, vaccination, or in the case of infants passively transmitted via maternal breastmilk, as well as the host microbiome. Examples of these sources of variability are summarized graphically in Figure 1.

1.3 | Contributions of This Paper

We make the following two arguments. First, as incidence is fundamentally about describing infection rates in populations or sub-populations, it fails to address the mechanisms by which individuals become infected. In incidence models which examine associations with risk factors, only the first source of variability listed above, that of the risk factors, is addressed. In the

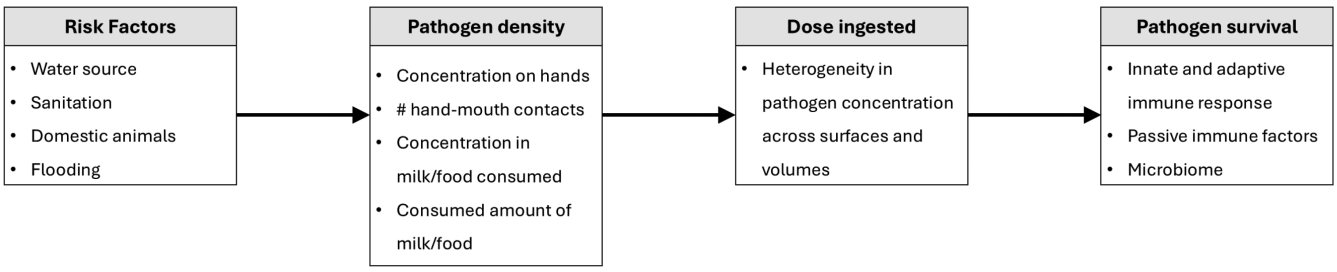


FIGURE 1 | Sources of variability in potential infections, along with examples of influential factors for each.

language of Haas et al. [10], such incidence models fail to achieve biological plausibility as an infection model. If the mechanisms of infection are the quantities of interest, then dose accrual rates, rather than incidence rates, ought to be the focus of one's analysis.

Second, stating that a unit increase in x leads to a η increase in the incidence rate is a fundamentally weaker statement than stating that a unit increase in x leads to a η increase in the pathogenic dose ingested for a given dose–response model. This claim is based on the fact that having knowledge about the effect a risk factor has on the dose accrual rate allows one to determine the effect on the incidence rate, whereas the reverse is not true.

To address these two points, we propose a novel model for assessing the impact of individual-level features on the dose accrual rates for the pathogen under study. Our proposed approach, which we call Dose Accrual Risk Estimation, or *DARE*, has the following features.

- By building on Quantitative Microbial Risk Assessment (QMRA) techniques, we both satisfy the plausibility criteria given by Haas et al. [10] and account for the salient sources of variability listed above.
- Rather than focusing on incidence estimation, our approach directly estimates the rate at which pathogenic dose is accrued per time unit.
- Our approach can handle varying time intervals between individuals' repeated measurements.
- We further provide a method for leveraging information across the analyses of multiple pathogens that uses recent work on linear subspace shrinkage techniques.
- Our DARE methodology is available through the R package *dare*, available through github.

The remainder of this paper is as follows. Section 2 describes our proposed longitudinal model with its derivation, along with an approach for leveraging information across the analyses of multiple pathogens. Section 3 describes a simulation study analyzing the estimation performance of our approach. Section 4 illustrates our proposed approach based on a subset of data collected through the PATHOME study [11]. Finally, we provide a discussion in Section 5.

2 | Methods

2.1 | A Dose Accrual Model

Suppose our study involves N at-risk individuals, and for the ith individual we observe them at the end of each of T_i intervals before either an infection is detected or they exit the study. The study design dictates the maximum value for T_i, and should an infection be detected, T_i will be less than or equal to the maximum allowed value. We will denote the length of these time intervals as τ_{it} and the binary outcome as y_{it}, where y_{it} equals one if after a period of τ_{it} they are infected and zero otherwise. For an infection to occur during an exposure window, a subject must ingest one or more infection-causing pathogens. In addition, for an infection to occur, one or more of these ingested pathogens must survive within the host long enough to begin the infection. Based on variations of these two processes, a plethora of dose–response models have been developed (see, e.g., Haas et al. [10]). Our proposed approach is agnostic to the specific dose–response model, in that what we propose should be compatible with any such model. We will therefore denote the dose–response model as $P_\theta(\cdot)$, where θ is the set of associated dose–response parameters.

The two most common dose–response models are the exponential model and the beta-Poisson model [12]. Note that while we focus on the most commonly used models, many others may be considered here as well. See, for example, Namata et al. [13] for descriptions of other such models. The only parameter of the exponential dose–response model is the (iid) survival rate of the organisms, and the model itself is given by

$$P_\theta(D) := 1 - e^{-\theta D}, \tag{1}$$

where D is the expected dose. The beta-Poisson model is predicated on host variability leading to different survival rates. The beta-Poisson model is then parameterized by, as its name suggests, the two shape parameters of the beta distribution describing the between-host distribution of organism survival probabilities. Thanks to theoretical work in Furumoto and Mickey [14], the beta-Poisson dose–response model is commonly approximated as

$$P_\theta(D) := 1 - \left(1 + \frac{D}{\theta_2}\right)^{-\theta_1} \tag{2}$$

If we were to know the mean dose D_{it} of the tth exposure period for subject i, the conditional likelihood of our data would be given as

$$\begin{aligned}&\Pr\left(y_{1,1}, \ldots, y_{N,T_N} \mid D_{1,1}, \ldots, D_{N,T_N}, \theta\right)\\ &= \prod_{i=1}^{N}\prod_{t=1}^{T_i}\left[P_\theta\left(D_{it}\right)\right]^{y_{it}}\left[1 - P_\theta\left(D_{it}\right)\right]^{1-y_{it}}.\end{aligned} \tag{3}$$

In highly controlled experiments, D_{it} would typically represent the dose concentration in a given medium being ingested. In observational studies, however, D_{it} in Equation (3) represents an agglomeration of pathways. That is, D_{it} can be thought of as an *expected ingested dose* that encompasses the collection of dose concentration on hands, in water, in milk, in food, and so forth, as well as behaviors that lead to ingestion such as hand-to-mouth contacts and quantity of media consumed. D_{it} should be considered stochastic for dose densities and time interval-specific behaviors will not be the same for two individuals with the same risk factors, nor even the same individual at different times.

Historically, expected pathogen dose has been modeled using log-normal distributions [10], and we do not break from this tradition here. Letting X_{it} denote a $1 \times J$ vector of risk factors for individual i during the tth time interval, the expected ingested dose, D_{it}, is a stochastic quantity depending on X_{it} (while the actual dose ingested is a separate stochastic quantity captured

in the dose–response model P_θ). Specifically, we model D_{it} as follows:

$$D_{it} \sim \ell N\left(X_{it}\boldsymbol{\beta} + \log(\tau_{it}), \sigma^2\right) \tag{4}$$

where $\ell N\left(\mu, \sigma^2\right)$ is the log-normal distribution with location parameter μ and log-scale parameter σ, and $\beta := \left(\beta_1, \ldots, \beta_J\right)$ is the vector of log-rate regression coefficients. Equation (4) implies that the rate of accrual of expected dose is

$$e^{X'_{it}\beta + \frac{\sigma^2}{2}} \tag{5}$$

leading to the first and second central moments of the expected dose over a time interval of length τ_{it} to be

$$\mathbb{E}\left(D_{it}|X_{it}, \beta\right) = \tau_{it}e^{X'_{it}\beta + \frac{\sigma^2}{2}}, \qquad \mathrm{Var}\left(D_{it}|X_{it}, \beta\right) = \mathbb{E}^2\left(D_{it}|X_{it}, \beta\right)\left(e^{\sigma^2} - 1\right)$$

One note of interest is that if the exponential dose–response model is used, as $\sigma \to 0$ we obtain the GLM based on the binomial distribution with the complementary log–log link function.

In summary, the two dose–response models described above assume that the number of pathogens ingested follows a Poisson distribution given the mean dose. Each pathogen ingested is assumed to either have a constant survival probability (exponential dose–response model) or host-specific survival probability (beta-Poisson dose–response model). The expected ingested dose itself, due to variability in the environment and host behavior, follows a log-normal distribution, and this distribution depends on observable risk factors. In particular, the rate at which dose is accrued changes by a factor of e^{β_j} from a unit increase in the jth covariate, keeping all other covariates the same. Together, we obtain the unconditional likelihood of our data as

$$\begin{aligned}&\Pr\left(y_{1,1}, \ldots, y_{N,T_N}|\theta, \beta, \sigma^2\right)\\ &= \prod_{i=1}^{N}\prod_{t=1}^{T_i}\int_{-\infty}^{\infty}\left[P_\theta\left(\tau_{it}e^{X'_{it}\beta + \sigma z_{it}}\right)\right]^{y_{it}}\left[1 - P_\theta\left(\tau_{it}e^{X'_{it}\beta + \sigma z_{it}}\right)\right]^{1-y_{it}}\phi\left(z_{it}\right)dz_{it}\end{aligned} \tag{6}$$

where $\phi(\cdot)$ is the standard normal probability density function. We will refer to Equation (6) as the DARE model. Note that in computing the DARE likelihood, the univariate integrals can be solved using standard numerical integration methods, such as Gaussian quadrature.

In the DARE model, there is an important issue of identifiability that must be acknowledged. In both the exponential and beta-Poisson dose–response models we have perfect confounding involving the intercept term. That is, β_1 (assuming $X_{it1} = 1 \forall i, t$) is perfectly confounded with θ from the exponential model and θ_2 in the beta-Poisson model. We therefore fix the θ or θ_2 to be 1, and estimate β_1 as an unconstrained (and uninterpretable) parameter. However, this issue further necessitates some caution in interpretation of any modeling results. Were we to have modeled $\log(\theta)$ in the exponential model or $-\log\left(\theta_2\right)$ in the beta-Poisson model[1] as a linear combination of our covariate vector X_{it}, the corresponding regression coefficients would again be perfectly confounded with β. As an anonymous reviewer pointed out, this may be ameliorated by strong prior information on model parameters for the dose–response model component and the dose accrual rate model component. In general, however, as one interprets any analysis output using the DARE model, one is bound to determine using context and domain expertise whether the effect of a specific covariate is on the dose accrual rate or the survival rate of the pathogens.

2.2 | Combining Results From Multiple Pathogens

We now expand our discussion to include contexts where we are measuring multiple pathogens. It will often be the case that a covariate will act on the rate of dose accrual similarly between certain pathogens if, for example, two pathogens are often transmitted through the same vector, fomite, or vehicle. Yet a hard constraint setting these regression coefficients to be equal is highly implausible. For example, if one pathogen is solely waterborne, while another pathogen is both waterborne and transmitted through food, both of these pathogens' rates of accrual will change similarly with respect to safe water access, yet clearly there will still be important differences; the former pathogen will be successfully mitigated through a water intervention, while the latter has a minimum threshold of effect that cannot reach disease elimination solely through such a water intervention. In other words, we wish to shrink certain regression coefficients toward each other in a data-driven way without imposing unrealistic hard constraints of equality. The recent SUBSET method of Sewell [15] provides tools to accomplish this. The idea is to find a linear subspace to shrink toward and use exponential tilting of the prior to induce the desired shrinkage. Unlike most statistical shrinkage methods which focus on point estimation, this approach shrinks the entire posterior, thereby influencing all resulting inference. By adapting SUBSET to our present context as described below, we are able to leverage information across pathogen-specific analyses in a data-driven way that, while not imposing any equality constraints, allows the data to dictate the degree to which certain parameters ought to be similar across pathogens.

Let $\beta_{(k)j}$ denote the jth regression coefficient for pathogen k, $k = 1, \ldots, K$, and similarly let $\theta_{(k)}$ and $\sigma^2_{(k)}$ be the dose–response model parameter(s) and dose variance for the kth pathogen respectively. The entire parameter vector of regression coefficients, dose response parameters, and dose variance is of length $Q := KJ'$, where $J' := (J + |\theta| + 1)$, and we will denote it as

$$\eta := \left(\beta_{(1)1}, \ldots, \beta_{(1)J}, \theta_{(1)}, \sigma^2_{(1)}, \beta_{(2)1}, \ldots, \beta_{(K)J}, \theta_{(K)}, \sigma^2_{(K)}\right)$$

Let L denote a matrix whose Q rows represent the unknown parameters and whose columns dictate which parameters are free and which have an equality constraint. We can construct L in the following manner, assuming that the K pathogens' intercept, θ, and σ^2 will not be shrunk toward each other. For a set $S \subseteq [K]$, let $v_M(S)$ denote the $M \times 1$ column vector such that its mth element equals 1 if $m \in S$ and 0 otherwise; and let $I(j, S)$ denote the $Q \times |S|$ matrix equal to the $Q \times Q$ identity matrix, selecting those $|S|$ columns corresponding to $\left\{J'(k-1) + j : k \in S\right\}$. Finally, for $j = 1, \ldots, J$ let $S_j \subseteq [1 : K]$ denote the set of pathogens whose

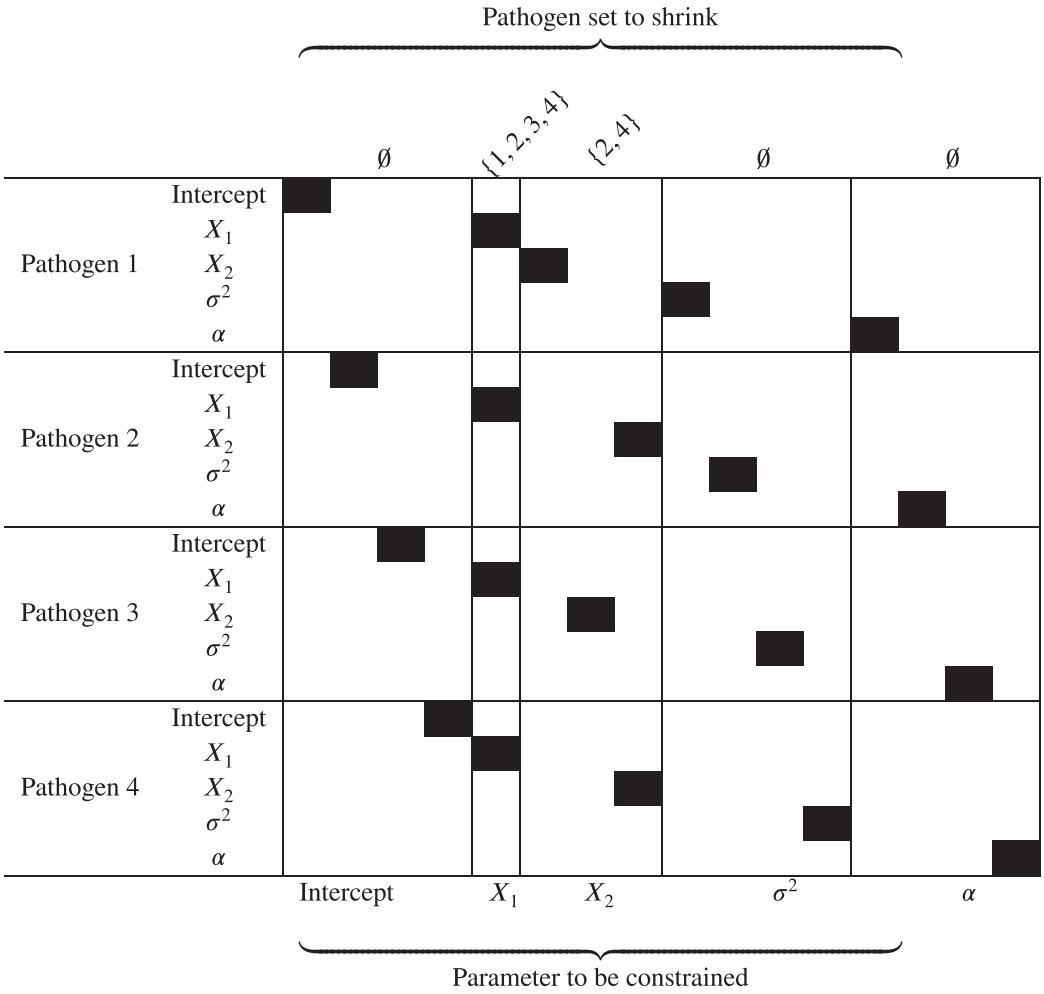


FIGURE 2 | Illustration of the matrix L to shrink certain parameters toward equality for certain pathogens. The intercept, σ^2, and α are not shrunk; the coefficients for X_1 are shrunk toward each other; and the coefficients for X_2 are shrunk toward each other for pathogens 2 and 4 only. The subsets S_j are labeled above, while the parameters being constrained are labeled below.

jth coefficients we wish to shrink toward each other, and S_j^C denote its complement.[2] Then we can set L to be

$$L \coloneqq \left(I(1,[K]), I\left(2, S_2^C\right), v_K\left(S_2\right) \otimes v_{J'}(\{2\}),\right.$$
$$\left. \ldots, I\left(J, S_J^C\right), v_K\left(S_J\right) \otimes v_{J'}(J), I(J+1,[K]), \ldots, I\left(J',[K]\right)\right) \quad (7)$$

The linear subspace we wish to shrink toward is span(L). Only the regression coefficients that have a $S_j \neq \emptyset$ experience shrinkage.

As an example, consider the matrix image of L given in Figure 2, where there are four pathogens; three covariates consisting of the intercept, X_1, and X_2; the intercepts are not shrunk toward each other; all X_1 coefficients are shrunk toward each other; and the X_2 coefficients are shrunk together for only pathogens 2 and 4. That is, $S_1 = \emptyset$, $S_2 = \{1,2,3,4\}$, $S_3 = \{2,4\}$ (and σ^2 and α do not experience shrinkage).

The SUBSET prior multiplicatively changes the joint prior on η by a factor of

$$\exp\left\{-\frac{\nu}{2}\eta'\left(I_Q - L\left(L'L\right)^{-1}L'\right)\eta\right\}, \quad (8)$$

which effectively penalizes areas of the parameter space distant from the linear subspace, where I_Q is the $Q \times Q$ identity matrix, and ν is a positive valued scalar that determines the level of shrinkage. The value of ν can be selected in a data-driven way by maximizing the Bayes factor (see [15], for details). If after analyzing the pathogens separately we denote the mode and the hessian of the negative log posterior for η as m_n and Ω_n respectively, then the large sample approximation of the joint posterior of η under the SUBSET prior induced by L is given by

$$\eta|\text{data} \sim N\left(\tilde{m}_n, \tilde{\Omega}_n^{-1}\right),$$
$$\text{where } \tilde{\Omega}_n \coloneqq \Omega_n + \nu\left(I_Q - L\left(L'L\right)^{-1}L'\right),$$
$$\tilde{m}_n \coloneqq \tilde{\Omega}_n^{-1}\Omega_n m_n \quad (9)$$

3 | Simulation Study

We wished to see how well we could estimate the unknown regression coefficients of the DARE model using the beta-Poisson

dose–response model, as well as determine how well the complementary log–log binomial GLM- the closest existing model to DARE- can recover the true parameter values. The true values of β in our simulation study were (−4.6,0,0.5,1) corresponding to an intercept and three covariates, each of which were randomly drawn for each individual from a standard normal distribution. We generated data according to both the exponential and the beta-Poisson. We considered all combinations of $\sigma \in \{1,2,3\}$ and- for the beta-Poisson model- $\theta_1 \in \{1,2,3\}$. These values, excluding the non-intercept regression coefficients, were selected to resemble the average values of those estimated from the PATHOME data described in Section 4, for which the average intercept (taken over all pathogens and age groups) was −4.6, the average σ was 3.0, and the average θ_2 was 2.5.

To evaluate our approach, we computed the coverage rates of the 95% credible intervals and examined the estimated regression coefficients. For each dose–response model, we simulated 1000 data sets. Each data set included 215 subjects, each measured at times 1, 3, 5, 7, and 14 or until an infection was detected; again, these values were selected to mimic the PATHOME study described in Section 4.

For both the DARE model and the complementary log–log GLM, we used weakly regularizing priors in order to provide numerical stability without unduly influencing the posterior. Specifically, we used a $N\left(0,2.5^2\right)$ prior for the regression coefficients with the exception of the intercept for which we used a $N\left(0,10^2\right)$ prior. For the DARE model fits, we used a gamma prior with shape and rate both equal to 2 for σ and an exponential with mean 1 for θ_1.

Figure 3 shows the coverage rates for the GLM as well as the DARE model using 95% central credible intervals. When $\sigma = 1$, the GLM has lower than nominal but respectable coverage rate. However, for larger values of σ the coverage becomes unacceptably low. As θ_1 increases, the coverage rate decreases, but not as severely as with increases in σ. Meanwhile, the beta-Poisson DARE model appears to maintain the nominal coverage rate, even when the dose–response model is misspecified. However, this comes at the expected cost of increased CI widths, as seen in Figure 4.

Figure 5 graphically displays the estimated regression coefficients from both the DARE and GLM model fits. From this we see that the GLM estimates tend to be negatively biased for non-zero coefficients while the DARE estimates tend toward a positive bias. However, while the DARE bias appears relatively stable across values of σ and θ_1, the negative bias of the GLM gets progressively worse as σ increases.

To determine the effect of model-misspecification on the expected dose, we conducted an additional, smaller, simulation study very similar to the one above with the following exceptions. First and foremost, we have replaced the error distribution on $\log\left(D_i\right)$ from normal (see Equation 4) to a t distribution with 5 degrees of freedom as well as to a zero-mean skew-normal distribution with tilting parameter equal to 5. Second, for simplicity we limited the true σ to equal 2 and the dose–response model to be beta-Poisson with $\theta_1 = 2$. The results matched closely with those results described above. The coverage rates were 0.97 and 0.94 for the t and SN distributions respectively, with average CI widths of 1.1 and 0.9 respectively. The estimates themselves seem to show the same overall patterns as shown in Figure 5, including the upward bias for the non-zero coefficients β_2 and β_3.

4 | Enteric Infections in Infants

The Pathogen Transmission and Health Outcome Models of Enteric Disease (PATHOME) study aims to use a One Health approach to better understand enteric pathogen transmission in low- to middle-income countries. Infants from 0 to 12 months old and their households were recruited into the PATHOME study from low-income and middle-income neighborhoods in Nairobi and Kisumu, Kenya. In each city, we selected households from both low and middle-income neighborhoods. At

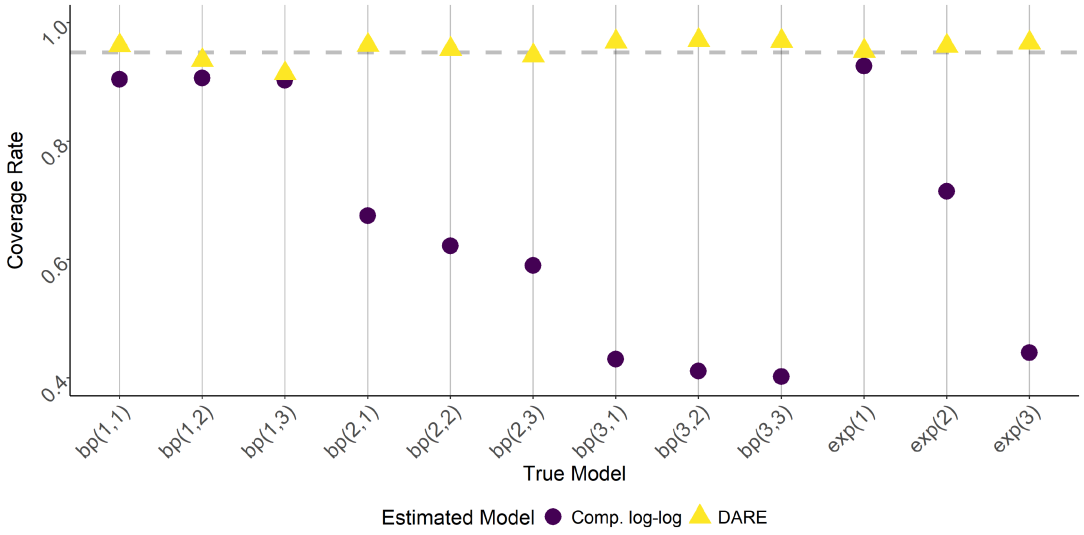


FIGURE 3 | Simulation study results for the coverage rate of 95% credible intervals aggregating over β_1, β_2, and β_3, comparing a GLM with the complementary log–log link with DARE based on the beta-Poisson dose–response model. The true model is given in the form $\mathrm{bp}\left(\sigma,\theta_1\right)$ or $\exp(\sigma)$ for the beta-Poisson and exponential dose–response models respectively. The nominal rate (0.95) is given in the dashed gray line.

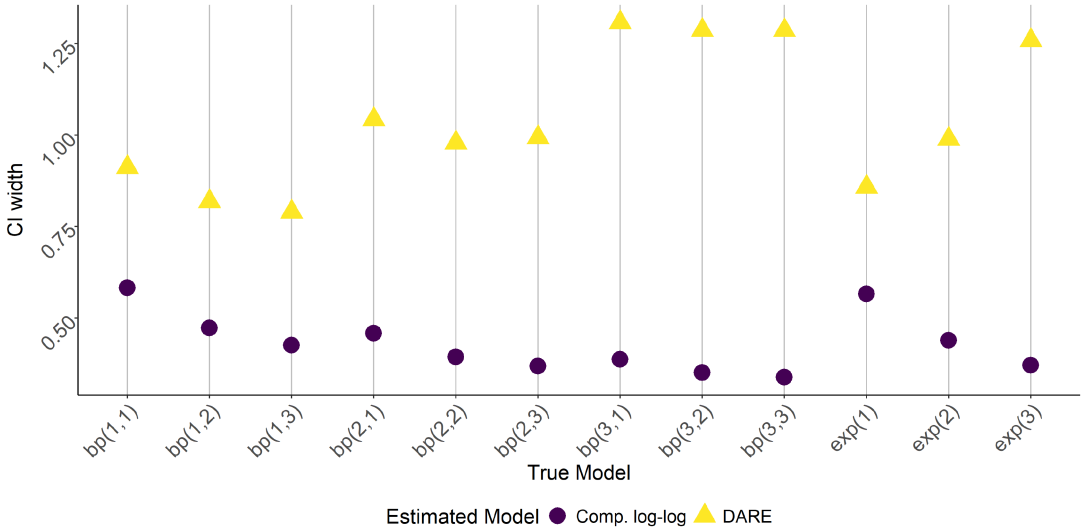
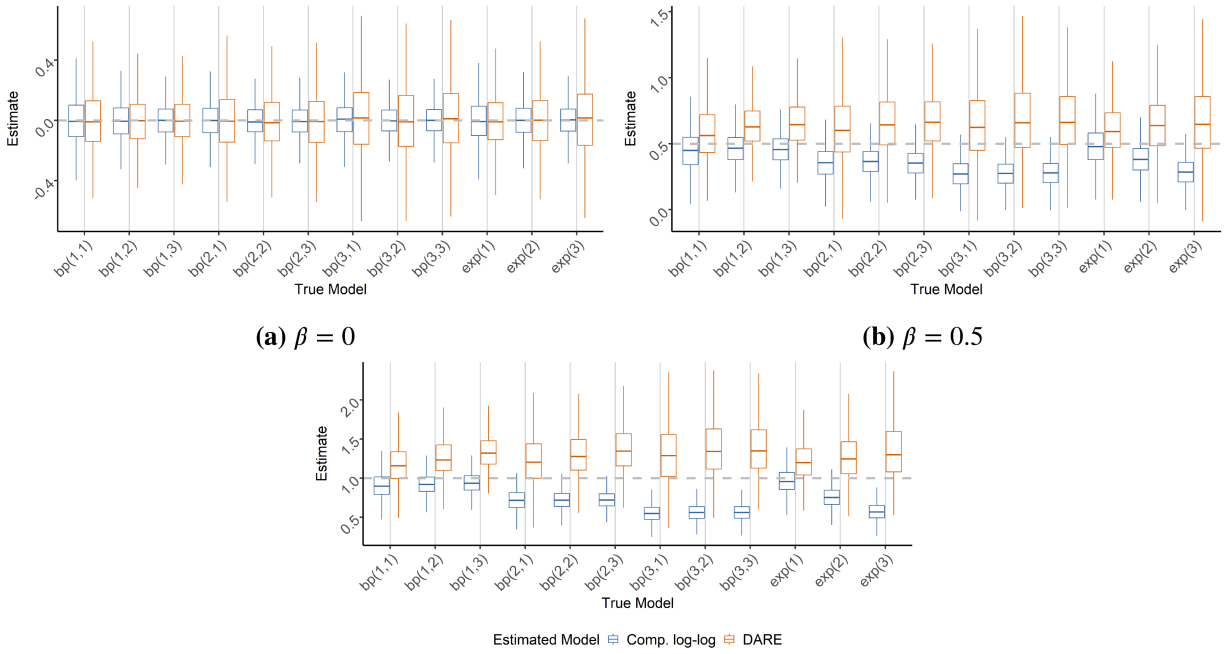


FIGURE 4 | Simulation study results for the central credible widths of 95% average credible intervals aggregating over β_1, β_2, and β_3, comparing a GLM with the complementary log–log link with DARE based on the beta-Poisson dose–response model. The true model is given in the form $\mathrm{bp}(\sigma, \theta_1)$ or $\exp(\sigma)$ for the beta-Poisson and exponential dose–response models respectively.



(a) $\beta = 0$ **(b)** $\beta = 0.5$

(c) $\beta = 1.0$

FIGURE 5 | Simulation study results for the point estimates of the three regression coefficients ($\beta_1 = 0$, $\beta_2 = 0.5$, and $\beta_3 = 1$), comparing a GLM with the complementary log–log link with DARE based on the beta-Poisson dose–response model. The true model is given in the form $\mathrm{bp}(\sigma, \theta_1)$ or $\exp(\sigma)$ for the beta-Poisson and exponential dose–response models respectively. The true values are given by the dashed gray lines.

the time of this study, we obtained microbiological data on 214 infants.

On the first day of participating in the study, caregivers were given a survey on socioeconomic conditions, behaviors, and household health. The variables we analyzed included city, socioeconomic status of the neighborhood (SES), whether the household's compound flooded, whether the household owned domestic animals,[3] whether animals not owned by the household entered their compound, and whether the family had access to a private latrine. These data elements are summarized in Table 1.

For each infant, diapers were provided to the caregivers in order to later collect stool samples on days 1, 3, 5, 7, and 14. If a stool was unavailable on a prespecified day, field staff would return the subsequent day to attempt diaper collection. The average (standard deviation) number of stool samples collected per child was 4.1 (1.2), and the counts of diapers collected on each day since

TABLE 1 | Summary statistics for infants enrolled in the PATHOME study.

Age	(0,90]	(90,180]	(180,270]	(270,360]
Number of infants	47	51	51	65
City (%)				
Kisumu	21 (44.7)	20 (39.2)	19 (37.3)	22 (33.8)
Nairobi	26 (55.3)	31 (60.8)	32 (62.7)	43 (66.2)
SES (%)				
Lower class	24 (51.1)	21 (41.2)	24 (47.1)	35 (53.8)
Middle class	23 (48.9)	30 (58.8)	27 (52.9)	30 (46.2)
Flood (%)				
No	42 (89.4)	43 (84.3)	42 (82.4)	56 (86.2)
Yes	5 (10.6)	8 (15.7)	9 (17.6)	9 (13.8)
Household owns animals (%)				
No	42 (89.4)	41 (80.4)	45 (88.2)	54 (83.1)
Yes	5 (10.6)	10 (19.6)	6 (11.8)	11 (16.9)
Neighborhood animals enter compound (%)				
No	33 (70.2)	33 (64.7)	36 (70.6)	49 (75.4)
Yes	14 (29.8)	18 (35.3)	15 (29.4)	16 (24.6)
Latrine (%)				
Private	35 (74.5)	40 (78.4)	37 (72.5)	48 (73.8)
Public	12 (25.5)	11 (21.6)	14 (27.5)	17 (26.2)

enrollment are given in Figure 6, stressing the importance of methods that can accommodate varying time lags between observations. Each stool sample was analyzed by quantitative molecular detection methods targeting unique pathogen-specific gene sequences. While pathogen presence/absence was assigned for 19 pathogens, most were too sparse to be used in our analyses given our sample size of 214 households at the time of this work. The pathogens included in this analysis were Enteroaggregative *E. coli* (EAEC), Enterotoxigenic *E. coli* (ETEC), typical enteropathogenic *E. coli* (tEPEC), atypical enteropathogenic *E. coli* (aEPEC), Shiga producing *E. coli* (STEC), *Campylobacter jejuni* (*C. jejuni*), *Salmonella*, and *Shigella*. The numbers of infections present at baseline, new infections (as defined to be an absence at baseline with a positive detection at some later follow-up), and no infections during the study interval are given in Table 2.

We fit the DARE model to each of the eight pathogens, and subsequently applied the SUBSET method described in Section 2.2 to leverage information across pathogens, selecting the amount of shrinkage based on Bayes factors. We limited shrinkage of the regression coefficients relating to animals to those for which animals have been shown to act as a vector, namely *aEPEC* (e.g., [16]), *STEC* (e.g., [17]), *ETEC* (e.g., [18]), *C. jejuni* (e.g., [19]), and *Salmonella* (e.g., [20]). As infants of various ages are likely to experience different exposures, we disaggregated infants into 3-month age groups: 0–3, 4–6, 7–9, and 10–12 months of age.

To assess the goodness-of-fit of the DARE model based on the beta-Poisson dose–response submodel, we computed the Bayesian posterior predictive p-value using the χ^2 discrepancy as a test statistic [21] for each pathogen and each age group. All p-values were between 0.20 and 0.73, indicating a good fit of the DARE model to the PATHOME data (values between 0.05 and 0.95 are typically considered to indicate reasonable fits of the data [22]).

Figure 7 shows the point estimates and credible intervals for the dose accrual rate ratios. Only animal ownership displayed statistical significance, and interestingly this occurred in 7–9 month-aged infants, when infants typically begin to crawl, but this relationship seemed to disappear in 10–12 month infants. For *ETEC*, *aEPEC*, *STEC*, *C. jejuni*, and *Salmonella*, we estimated the dose accrual rate to be 5.3, 4.6, 6.0, 5.0, and 5.0 times higher, respectively, for those whose households owned animals compared to those whose households didn't. The posterior probabilities that these rate ratios were greater than one exceeded 0.99 for these five pathogens.

We also examined the effect of animal ownership on incidence proportion for these five pathogens, holding all other covariates fixed at their modal value; these values are given in Figure 8. For *ETEC*, *aEPEC*, *STEC*, *C. jejuni*, and *Salmonella* respectively, the incidence proportion rate for 7–9 month olds was estimated to be 3.2, 2.5, 3.3, 2.5, and 2.6 times higher for those whose households owned animals compared to those whose households did not. It is sensible that animal ownership was not significant for younger age groups as infants begin to learn to crawl around 7 months of age [23]. This suddenly gives them much more access to animal feces on the ground as well as soil, both indoors and outdoors in

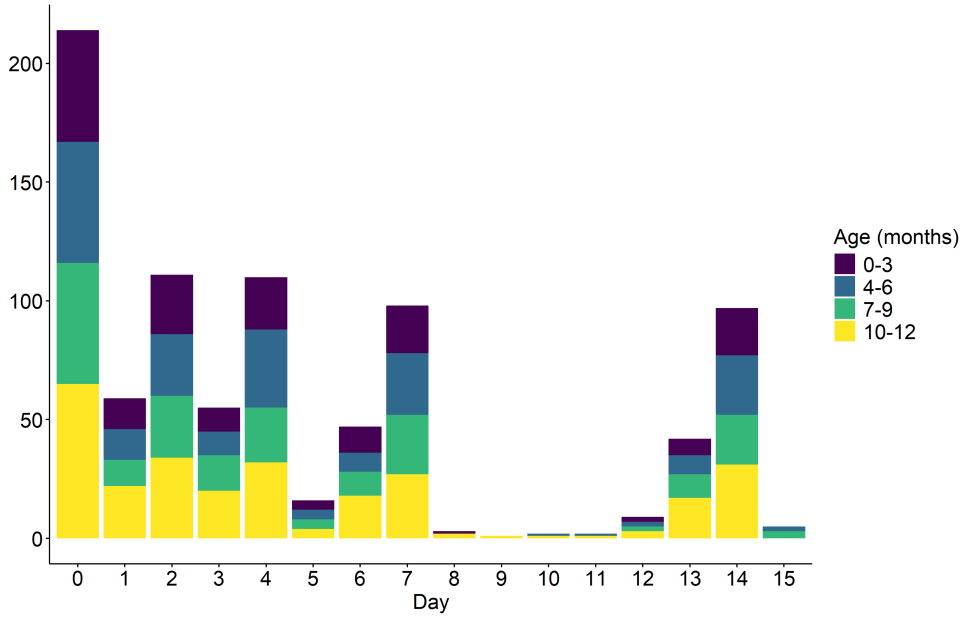


FIGURE 6 | Number of infant diapers collected by day, broken down by age.

TABLE 2 | Pathogen detection summary for infants enrolled in the PATHOME study.

Pathogen	Present at baseline	New infection	No infection
EAEC	107	54	53
ETEC	35	33	146
tEPEC	14	19	181
aEPEC	41	38	135
STEC	15	27	172
Campylobacter jejuni	18	16	180
Salmonella	39	37	138
Shigella	60	29	125

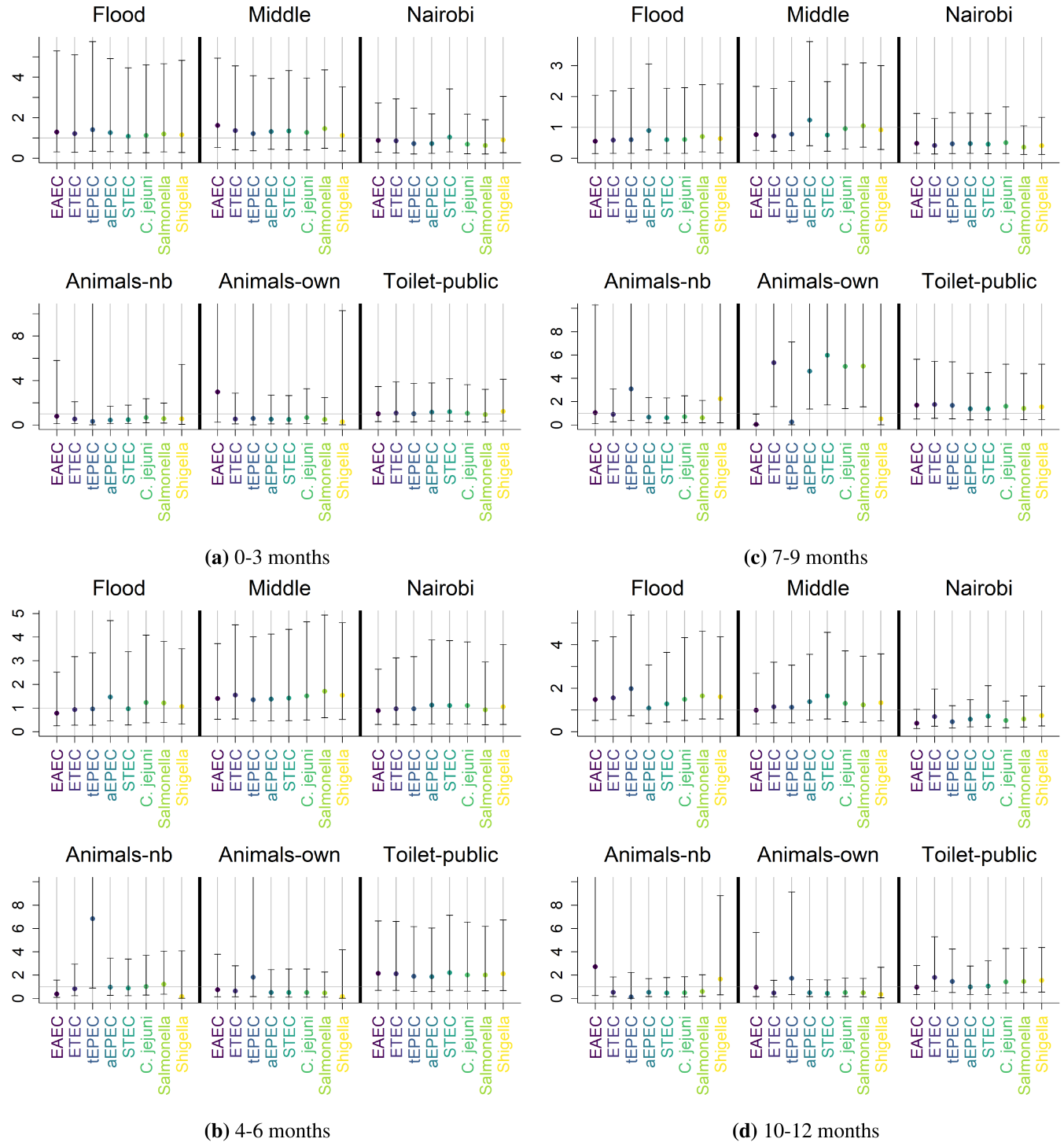
the compound, which can be contaminated by animal feces; in ongoing work we have found roughly half of soil samples taken from these PATHOME households (regardless of SES) were positive for *E. coli*, roughly half were positive for Salmonella, and roughly 10% were positive for Shigella. This is also consistent with observed infant behaviors described by Tumwebaze et al. [24], showing that the odds of contacting animal feces was ≈2.5 times greater for 7–10 month olds than 0–3 month olds and ≈4 times greater odds of being exposed to surfaces contaminated with animal feces. The effect of animal ownership was not significant for the 10–12 month age group, and this might be because of some degree of immunity built up during the prior months of life or because it is simply a false negative result.

5 | Discussion

Enteric disease is a significant source of morbidity and mortality globally, especially in children living in low- to middle-income countries. Enteric diseases occur through the ingestion of pathogens, and understanding how various risk factors affect dose accrual rates is vital to developing effective interventions. We have proposed a novel approach to estimating the effect individual-level characteristics have on pathogenic dose accrual rates. Our approach provides a biologically plausible infection model that attempts to account for the four sources of variability we have described- risk factors, dose concentration, number of pathogens ingested, and pathogen survival rate (see Figure 1). In contrast to incidence models, our approach focuses on the mechanisms of infections, and by modeling the dose accrual rates we can make stronger statements than simply estimating incidence rates.

Our approach allows measurements taken at unevenly spaced time points, can flexibly handle various disease-appropriate dose–response models, and our simulation study suggests that the coverage rate is maintained at or very nearly at the nominal rate even under misspecification of the dose–response model. We have further provided a method for leveraging information across multiple pathogens within a single study. The priors described in the simulation study were the same used in the PATHOME data analyses. However, more informative priors can and should be used should such information exist from prior studies or domain expertise.

We anticipate in most cases it will be clear if a risk factor may have an effect on dose accrual rates vs. within-host pathogen survival probability. In such cases, parameter interpretability of DARE is a strength, as the exponentiated regression coefficients provide the dose accrual rate ratio corresponding to a unit increase in the covariate of interest, holding all other covariates constant. However, interpretability is a limitation of our proposed approach in cases where the risk factor may affect both the dose accrual and the dose survival rate. There is perfect confounding in this situation. Intuitively this makes sense and appears to be unavoidable, since without exceedingly granular microbiological data collection procedures, it is not possible to disambiguate the number



(a) 0-3 months

(c) 7-9 months

(b) 4-6 months

(d) 10-12 months

FIGURE 7 | Dose accrual rate ratios for infants living in Kisumu and Nairobi, Kenya. The reference categories for "Middle", "Nairobi", and "Toilet-public" are low-class neighborhood, Kisumu, and access to a private toilet, respectively. "Animals-nb" refers to whether neighborhood animals enter the compound, and "Animals-own" refers to whether the household owns domestic animals.

of pathogens ingested from the pathogens' survival rates, when both survival and dose accrual depend on the same factor(s). It is, however, still possible to understand the general pattern (positive or negative) in the incidence rate due to such a risk factor, even if the precise cause is unknown.

Another potential limitation is the simplifying assumption that each individual's time intervals are independent. We feel this is typically a reasonable assumption because of two factors. First, correlation within an individual can typically be captured by measuring appropriate covariates. Second, there is an enormous amount of stochasticity in what pathogens are encountered and in what density, what behaviors and behavioral combinations occur (e.g., number of times touching a surface + mouthing hands), and so forth, and this variation ought to overwhelm any remaining individual-level correlation. Yet if this assumption does not hold, further extensions of the DARE model are required.

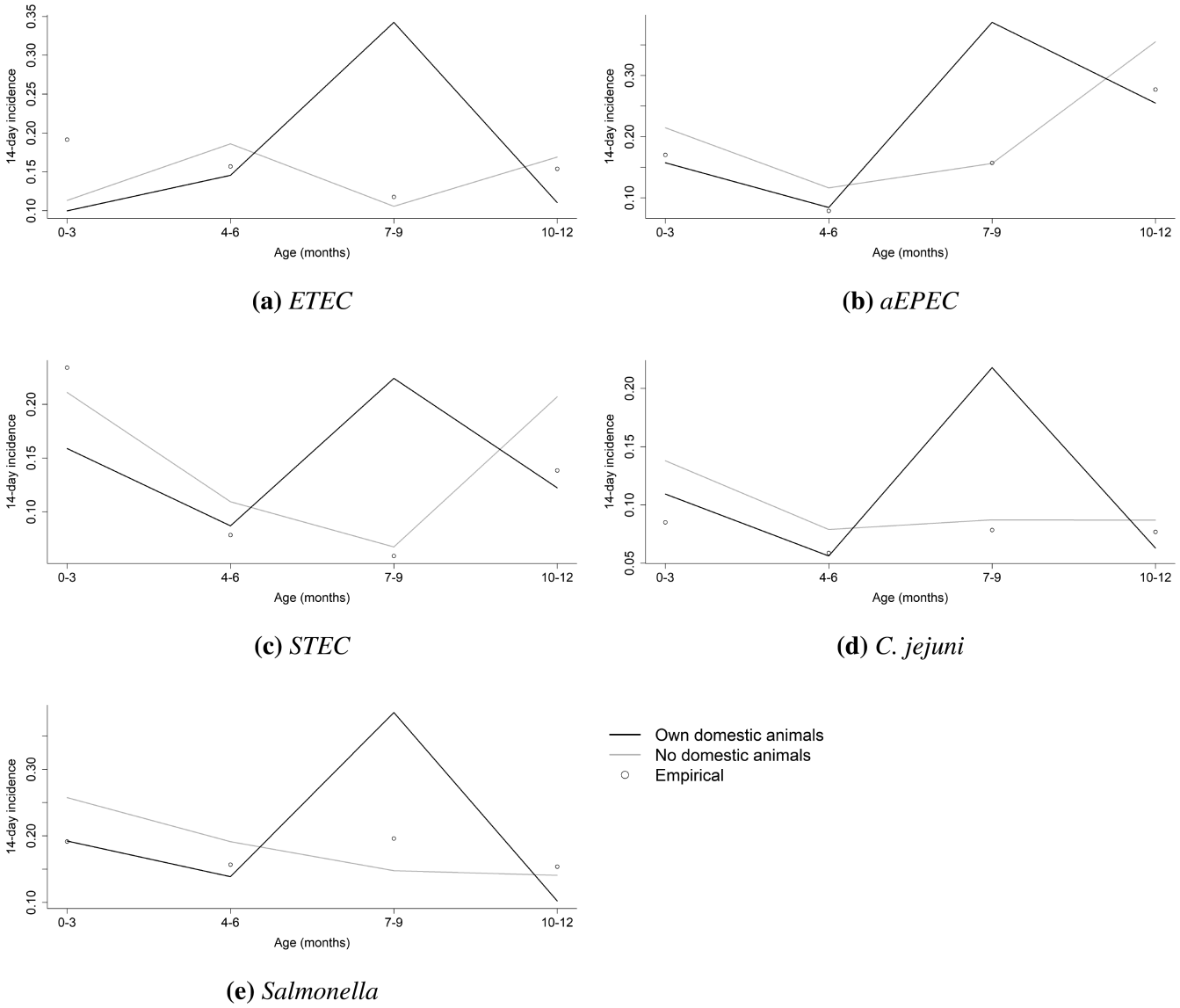


FIGURE 8 | Estimated 14-day incidence proportion by domestic animal ownership, holding all other covariates fixed at their modal value.

The DARE model provides an advancement over simpler incidence models, allowing a better and more nuanced understanding of how risk factors lead to higher infection rates. Our proposed methods can be implemented in the R programming language [25] via the R package found at https://github.com/dksewell/dare.

Acknowledgments

This work was funded by the National Institutes of Health Fogarty Institute Grant Number R01 TW011795.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The PATHOME data used in this work is not currently publicly available due to ongoing preparations for data sharing. However, the full PATHOME dataset is expected to be made publicly accessible within the near future. Once available, access details and relevant documentation will be provided via a publicly accessible repository, and the corresponding author can be contacted for updates regarding data availability.

Endnotes

[1] Note that as $-\log(\theta_2)$ increases, or equivalently θ_2 decreases, the survival probability of each organism increases.

[2] Note that S_j could be the empty set, in which case no shrinkage is performed on the jth regression coefficients.

[3] This included chickens, ducks, pigs, goats, sheep, cows, dogs, rabbits, donkeys, and turkeys.

References

1. World Health Organization, "Diarrhoeal Disease," (2024), https://www.who.int/news-room/fact-sheets/detail/diarrhoeal-disease.

2. P. Feuerstadt, N. Theriault, and G. Tillotson, "The Burden of CDI in the United States: A Multifactorial Challenge," *BMC Infectious Diseases* 23, no. 1 (2023): 132, https://doi.org/10.1186/s12879-023-08096-0.

3. M. M. Ward, "Estimating Disease Prevalence and Incidence Using Administrative Data: Some Assembly Required," *Journal of Rheumatology* Canada, volume 40 (2013): 1241–1243.

4. N. Bruce, D. Pope, and D. Stanistreet, *Quantitative Methods for Health Research: A Practical Interactive Guide to Epidemiology and Statistics* (Wiley, 2008).

5. E. L. Frome and H. Checkoway, "Use of Poisson Regression Models in Estimating Incidence Rates and Ratios," *American Journal of Epidemiology* 121, no. 2 (1985): 309–323, https://doi.org/10.1093/oxfordjournals.aje.a114001.

6. C. Cohen, J. Kleynhans, A. von Gottberg, et al., "Sars-Cov-2 Incidence, Transmission, and Reinfection in a Rural and an Urban Setting: Results of the Phirst-c Cohort Study, South Africa," *Lancet Infectious Diseases* 22, no. 6 (2022): 821–834, https://doi.org/10.1016/S1473-3099(22)00069-X.

7. N. G. Schwarz, A. A. Adegnika, L. P. Breitling, et al., "Placental Malaria Increases Malaria Risk in the First 30 Months of Life," *Clinical Infectious Diseases* 47, no. 8 (2008): 1017–1025, https://doi.org/10.1086/591968.

8. M. L. Verburgh, A. Boyd, F. W. N. M. Wit, et al., "Similar Risk of Sars-Cov-2 Infection and Similar Nucleocapsid Antibody Levels in People With Well-Controlled Hiv and a Comparable Cohort of People Without Hiv," *Journal of Infectious Diseases* 225, no. 11 (2021): 1937–1947.

9. A. Parker, M. A. E. Lawson, L. Vaux, and C. Pin, "Host-Microbe Interaction in the Gastrointestinal Tract," *Environmental Microbiology* 20 (2018): 2337–2353.

10. C. N. Haas, J. B. Rose, and C. P. Gerba, *Quantitative Microbial Risk Assessment* (John Wiley & Sons, 2014).

11. K. K. Baker, S. Simiyu, P. Busienei, et al., "Protocol for the Pathome Study: A Cohort Study on Urban Societal Development and the Ecology of Enteric Disease Transmission Among Infants, Domestic Animals and the Environment," *BMJ Open* 13 (2023): e076067.

12. J. A. Soller, "Use of Microbial Risk Assessment to Inform the National Estimate of Acute Gastrointestinal Illness Attributable to Microbes in Drinking Water," *Journal of Water and Health* 4, no. S2 (2006): 165–186.

13. H. Namata, M. Aerts, C. Faes, and P. Teunis, "Model Averaging in Microbial Risk Assessment Using Fractional Polynomials," *Risk Analysis* 28, no. 4 (2008): 891–905, https://doi.org/10.1111/j.1539-6924.2008.01063.x.

14. W. A. Furumoto and R. Mickey, "A Mathematical Model for the Infectivity-Dilution Curve of Tobacco Mosaic Virus: Theoretical Considerations," *Virology* 32, no. 2 (1967): 216–223, https://doi.org/10.1016/0042-6822(67)90271-1.

15. D. K. Sewell, "Posterior Shrinkage Towards Linear Subspaces," *Bayesian Analysis* 20 (2024): 657–680.

16. G. Krause, S. Zimmermann, and L. Beutin, "Investigation of Domestic Animals and Pets as a Reservoir for Intimin- (Eae) Gene Positive *Escherichia coli* Types," *Veterinary Microbiology* 106, no. 1 (2005): 87–95, https://doi.org/10.1016/j.vetmic.2004.11.012.

17. M. Čobeljić, B. Dimić, D. Opačić, Z. LEPŠANOVIĆ, V. Stojanović, and S. Lazić, "The Prevalence of Shiga Toxin-Producing *Escherichia coli* in Domestic Animals and Food in Serbia," *Epidemiology and Infection* 133, no. 2 (2005): 359–366, https://doi.org/10.1017/S0950268804003334.

18. D. J. Daniel, E. Isaacson Richard, and M. Schifferli Dieter, "Animal Enterotoxigenic *Escherichia coli*," *EcoSal Plus* 7, no. 1 (2016): 2016, https://doi.org/10.1128/ecosalplus.esp-0006-2016.

19. P. A. Manser and R. W. Dalziel, "A Survey of Campylobacter in Animals," *Journal of Hygiene* 95, no. 1 (1985): 15–21, https://doi.org/10.1017/S0022172400062239.

20. J. Oloya, M. Theis, D. Doetkott, N. Dyer, P. Gibbs, and M. L. Khaitsa, "Evaluation of Salmonella Occurrence in Domestic Animals and Humans in North Dakota (2000–2005)," *Foodborne Pathogens and Disease* 4, no. 4 (2007): 551–563, https://doi.org/10.1089/fpd.2007.0014.

21. A. Gelman, X.-L. Meng, and H. Stern, "Posterior Predictive Assessment of of Model Fitness via Realized Discrepancies," *Statistica Sinica* 6, no. 4 (1996): 733–760.

22. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, Third ed. (Chapman & Hall/CRC, 2004).

23. K. Butcher, A. VandenBerg, and T. Dodds, "Why Crawl?" (2013), https://www.canr.msu.edu/news/why_crawl.

24. I. K. Tumwebaze, M. Krysan, P. K. Busienei, et al., "Domestic Animals and Hygiene on Infants' Risk of Contact and Exposure to Animal Faeces in Urban Neighbourhoods in Kenya," (2025), https://ssrn.com/abstract=5334670, https://doi.org/10.2139/ssrn.5334670.

25. R Core Team, "R: A Language and Environment for Statistical Computing," (2023), R Foundation for Statistical Computing Vienna, Austria, https://www.R-project.org/.