©Copyright 2014
Rebecca Yates Coley

Bayesian Hierarchical Frailty Models for Heterogeneity in Risk

Rebecca Yates Coley

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Elizabeth R. Brown, Chair

James P. Hughes

Lurdes Y. T. Inoue

Megan Othus

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

Bayesian Hierarchical Frailty Models for Heterogeneity in Risk

Rebecca Yates Coley

Chair of the Supervisory Committee:
ScD Elizabeth R. Brown
Biostatistics

The effect of an intervention or exposure on time-to-event is most commonly estimated with the Cox model, which assumes proportional hazards. When heterogeneity in risk is present, the assumption of proportionality is violated and the Cox model's population-averaged estimate of the hazard ratio can underestimate the effect for an individual. In this dissertation, we develop Bayesian hierarchical frailty models that adjust for heterogeneity in risk in a way that reflects the sources of heterogeneity. We focus on the compound Poisson distribution, as it models a biological risk mechanism seen in many applications where an individual's risk of an event is the result of independent, competing exposure processes and it allows that some individuals have no risk of an event. In Chapter 2, we outline a hierarchical definition of the compound Poisson distribution and demonstrate Bayesian estimation of a hierarchical compound Poisson frailty model. In Chapter 3, we extend the model proposed in Chapter 2 to allow frailty distributions to vary across latent risk classes, resulting in a compound Poisson mixture frailty distribution. Risk-related covariates are used to classify participants into latent risk groups. We conclude with a discussion of future work, including ideas for a gamma frailty model with subject-specific parameters for situations in which exposure processes are observed.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to extend my deepest gratitude to my advisor, Elizabeth Brown, for demonstrating excellence in her own work and challenging me to do the same, and to my committee whose insight and encouragement have been invaluable. Thank you.

This work was supported by NIAID grant 5 UM1 AI068615.

DEDICATION

This dissertation is dedicated to my parents, David and Kenna Coley, for their limitless love and guidance. Thank you for supporting me when I decided to move across the country to take on the biggest challenge of my life, for understanding when I didn't call for days because I was busy working, and for instilling in me the dedication and passion needed to bring this project to completion.

Chapter 1

INTRODUCTION

1.1 Population- vs. Individual-Level Effectiveness

In studies where the outcome is time-to-event, the hazard ratio associated with an intervention or exposure is typically estimated with the Cox proportional hazards model [18]. Under the assumption that the effect of a covariate on the hazard is multiplicative, the Cox model produces a population-level estimate of the hazard ratio. When risk of an event varies across study participants, an intervention or exposure that acts multiplicatively on an individual's risk will not have a proportional impact on the population-level. Consequently a population-level estimate from the Cox model does not estimate the individual-level effect of an intervention or exposure when heterogeneity in risk is present. In this sense, the hazard ratio estimated by the Cox model is actually a biased estimate of the measure we may be most interested in, the reduction of risk for an individual receiving the intervention.

The impact of unobserved heterogeneity is most often discussed in the context of observational studies, but it also affects survival analysis of data from randomized trials. While randomization of treatment assignment at baseline offers some assurance that heterogeneity is balanced across study arms, the risk profiles of each arm do not necessarily remain comparable over the study duration. In the proportional hazards model, assessment of intervention effectiveness is made over time. When heterogeneity in risk is present, survival is compared across groups with different distributions of risk. As a result, the population-average effect of an intervention differs from the effect an intervention has on an individual level.

Consider a randomized trial for an intervention that reduces an individual's risk of an event. If heterogeneity in risk is present, risk of an event will vary within study arms, but the distribution of risk in each arm will be similar at baseline, due to randomization. As the study progresses, however, and events accrue, the distribution of risk in each arm will

not remain similar. Specifically, at a later time point, the intervention arm will contain participants with higher risk, on average, due to the intervention's protective effect. Participants in the control arm with comparably high risk will be more likely to have an event, or have an event sooner, as they are not benefitting from the intervention. As the proportion of individuals at higher risk of an event in the control arm is depleted, the intervention appears less effective. If this scenario is extended to include individuals with no risk of an event, a population-level analysis can find a truly effective intervention to be harmful. If the intervention is highly effective at prolonging survival, all at-risk participants in the control arm should experience an event before all those in the intervention arm do. After this point, incidence in the control group will necessarily be zero while incidence in the intervention arm remains positive such that the intervention appears to increase risk.

The degree to which Cox population-level estimates differ from the true effect of a covariate on an individual's hazard depends on the distribution of heterogeneity in the study population, as first explored by Aalen (1988). To illustrate this difference, we consider two models for heterogeneity: gamma and compound Poisson. The gamma distribution is the traditional choice for a *mixing distribution* modeling heterogeneity in risk. A compound Poisson random variable is equal to zero with some positive probability and otherwise distributed on the positive half-line. (We will discuss the construction of a compound Poisson random variable extensively in this dissertation but provide this example first without further details.) The key distinction between these two distributions is the presence of a point mass at zero, representing a subset of the population with no risk of an event. We calculated the population-averaged hazard ratio (HR), i.e. the expected HR estimate from the Cox model, for each mixing distribution using the formula for population intensity presented by Hougaard (1984) and used by Aalen (1988), which we detail in Appendix A.

Figure 1.1 shows the population-averaged HR over time when heterogeneity is gamma-distributed and compound Poisson-distributed for a covariate that reduces risk for an individual by 50%, i.e. HR=0.5. In both scenarios, population-averaged HRs are greater than 0.50 and closer to the null value of HR=1 (no effect), underestimating the covariate's effect for an individual. Estimate attenuation occurs more quickly when the variance of the mixing distribution is greater (i.e. there is increased heterogeneity). This discrepancy

is further pronounced when heterogeneity follows the compound Poisson distribution. In fact, when a larger proportion of the study population is not at risk of an event, we see that the population-averaged HR may actually be greater than one, indicating that the covariate increases risk. It can by shown that the limit (as time goes to infinity) of the population-averaged HR is one under a gamma mixing distribution and greater than one under a compound Poisson mixing distribution. (The exact limit for the compound Poisson depends on specific parameter values.)

1.2 Modeling Heterogeneity

There are several approaches that adjust for heterogeneity in risk in time-to-event models. The most common method is to include additional risk-related covariates in a Cox regression model. However, this approach is often insufficient for several reasons. First, complete and accurate data on all risk factors is often not available. Even if this information were available, the assumption of proportionality would be necessary for all covariates. Furthermore, in situations where some participants are not at risk, the model for heterogeneity should reflect this possibility. Yet, it would be difficult to constrain the Cox regression model so as to guarantee that near-zero risk would be predicted for some individuals. These limitations exist in both clinical trial and observational data settings.

Frailty and cure rate models offer alternative approaches to account for heterogeneity in risk. Consider the proportional hazards model:

$$\lambda(t)\exp(\mathbf{x}\boldsymbol{\beta})$$

where $\lambda(t)$ is a baseline hazard function, $\mathbf{x}$ is a vector of covariates and $\boldsymbol{\beta}$ is a vector of coefficients. Frailty modeling refers to the addition of a frailty term, denoted here with Z, to the hazard function in order to account for unobserved heterogeneity [50]:

$$\lambda(t)\, Z \exp(\mathbf{x}\boldsymbol{\beta}).$$

For identifiability, the expectation of frailty, $E(Z)$, is set equal to one. Hence, the frailty term is simply a risk modifier, scaling the hazard up or down, and the baseline hazard, $\lambda(t)$, can be interpreted as the average baseline hazard across the population. The survival

function under the proportional hazards model with frailty is:

$$S(t|x,Z) \quad = \quad \exp\Big(- Z \exp(x'\beta)\, \Lambda(t)\Big),$$

where $\Lambda(t) = \int_0^t \lambda(t)\, dt$ is the cumulative baseline hazard at time t.

Cure rate models are a method for modeling heterogeneity when a fraction of the population is not at risk of an event or is *cured* [10]. We assume that a proportion of the population, ϱ, is cured or not at risk of an event, while the remaining $1 - \varrho$ are non-cured and have some positive risk. The standard cure rate model is:

$$S(y) = \varrho + (1 - \varrho)S^*(y),$$

where $S^*(y)$ is the survival function for the at risk proportion of the population. Covariates can be included in S^* through a proportional hazards model. However, proportionality will not hold if these covariates are also used to model the cured fraction, ϱ, through logistic regression. It is also possible to incorporate frailty into $S^*(y)$.

Frailty modeling requires specification of a mixing, or frailty, distribution for Z. The gamma distribution is commonly used for modeling frailty, but this choice is typically made for the sake of mathematical convenience and not based on biological rational [26]. It is preferable, however, to select a frailty distribution that reflects the possible sources of heterogeneity in risk. Furthermore, when it is likely that some participants have no risk of an event, the model for heterogeneity should allow for that possibility, either through a cured fraction or implicit in the specified frailty distribution.

In Chapter 2, we consider the compound Poisson distribution for modeling frailty, which models an individual's risk of an event as the result of independent, competing exposure processes, a risk mechanism common in medical contexts. The use of the compound Poisson distribution as a frailty distribution was first suggested by Aalen (1988) and has been used in many applications [2, 3, 38, 39, 43, 45, 48]. We add to the existing literature by providing a hierarchical definition of the compound Poisson distribution, which highlights how the construction of compound Poisson frailty can reflect the risk mechanisms that result in heterogeneity. The number of exposure processes for each individual, frequently either measured with error or unknown, is modeled as a Poisson random variable, which allows

for a participant to have no risk (i.e. no exposure). Unlike a cure rate model, however, the parameter dictating the probability of risk for an individual, the mean parameter for the Poisson distribution, also influences the distribution of risk among those who are at risk. Each exposure process is assumed to present a some level of risk, which is modeled as a Gamma random variable. Under the compound Poisson model, separate exposure processes pose independent, competing risks for an individual; given exposure, one of the processes eventually leads to an event. An individual's risk of an event during a specified time period is the sum of these competing risks. A more technical definition of the compound Poisson random distribution is given in Chapter 2. This research is also the first to implement Bayesian estimation of a compound Poisson frailty model.

A limitation of current frailty methods for univariate survival data is that only information available in event time and study arm is used to model heterogeneity. While additional data collected on participants' risk factors could improve frailty estimation, existing methods for univariate survival data provide no mechanism for using risk-related covariates to inform individuals' frailties. With this motivation in mind, we have developed a mixture frailty model that uses risk-related covariate data to determine distinct frailty distributions. In Chapter 3, we propose a model that adjusts for heterogeneity in risk by varying frailty distribution parameters as a function of individual covariates, resulting in a mixture of frailty distributions. Ordinal latent class regression is used to identify groups of participants with similar risk characteristics. The compound Poisson distribution is used to model frailty for all classes. This distribution allows that some participants have no risk of an event and generally reflects how risk is accrued. Following our hierarchical definition of the compound Poisson distribution, risk-related covariates will be used to model two parameters: each class will have a different mean number of exposure processes and different shape parameter for the gamma random variable modeling risk associated with each process. The compound Poisson mixture frailty model invites comparison with the cure rate model, which can similarly model the probability of cure with risk-related covariates. However, the inclusion of frailty for those at risk in a cure-rate model is also subject to the limitations of frailty models for univariate survival data discussed above. In the proposed model, participant risk factors influence both the probability and level of exposure through the mean parameter of the

Poisson distribution of number of exposure processes, which reflects a natural expectation that characteristics that increase an individual's risk of exposure likely also increase the level of risk given exposure.

In Chapter 4, we outline ideas for a related model to account for heterogeneity in scenarios where the number of exposure processes is observed, instead of latent, but events are still observed on the individual-level. In previous chapters, the number of of exposure processes is not observed, so it is necessary to model the risk associated with each on an individual's exposure processes with the same distribution. In this context, with the number of exposure processes known, it is possible to model the risk associated with each exposure process separately. In this chapter, we propose a model that uses risk-related covariates measured at the exposure process-level, in addition to those measured at the individual-level, to model risk of an event. As a result, the effect of exposure process-level characteristics on an individual's risk can be assessed, in addition to estimation of an individual-level hazard ratio for the intervention or condition of interest.

While we arrive at this model from a competing risks framework, the method outlined in Chapter 4 addresses another problem in survival analysis, that of clustered time-to-event data where the event is observed on the cluster- (individual-) level while the unit (exposure process) experiencing the event is unknown. Existing methods for clustered time-to-event data assume that event time is observed for each unit within a cluster. Additionally, current frailty models for clustered time-to-event data do not accommodate unit-level covariate information. The model considered in Chapter 4 addresses these limitations.

1.3 Motivating Examples

1.3.1 HIV Prevention Trials

The possible discrepancy between intervention effectiveness for at-risk individuals and the Cox model population-level estimate of effectiveness is particularly concerning in the context of HIV prevention trials, where heterogeneity is expected. A study participant's risk of seroconversion depends on the number and type of sex acts with infected partners, infected partners' viral load, and other factors that will vary across individuals. Furthermore, in

general population studies, participants with no HIV-infected sexual partners during the study period are not at risk of seroconversion. Even in serodiscordant partner studies, a participant may have essentially no risk of seroconversion if their partner's viral load is very low or if they always use condoms.

Heterogeneity in risk offers a possible explanation as to why interventions that were highly effective in reducing risk of seroconversion in some recent trials have shown a negligible impact in others. While attention has recently turned to the possibility that antiretroviral therapy (ART), known to be effective in treating HIV, may also prevent HIV infection, studies of this approach have been inconclusive. Several trials have shown topical microbicides and oral chemoprophylaxis to be effective in reducing HIV transmission [4, 8, 47]. However, other trials of similar interventions have found no effect [5, 33, 35, 49]. These inconsistent results are often attributed to low adherence, but they may also be due to heterogeneity. Certainly, the presence of heterogeneity in a trial with poor adherence will lead to further underestimation of intervention effectiveness. We also note that heterogeneity may influence effectiveness estimates of interventions where adherence is not a concern, e.g. vaccines. In fact, heterogeneity in risk has been suggested as a possible reason for the apparent waning effectiveness of the RV144 vaccine [42].

Accounting for heterogeneity in risk in survival models used to analyze data from HIV prevention trials may enable better estimation of an intervention's effect on an exposed individual's risk of seroconversion. The various factors that influence an individual's risk of seroconversion are not observed during HIV prevention trials, so they cannot be included as covariates in a Cox regression model. Moreover, even if all factors were measured without error, inclusion in the Cox regression model would still require the assumption of proportional hazards. Using frailty methods to model unobserved heterogeneity is an appealing alternative in this context. Frailty modeling has previously been used in the analysis of HIV vaccine efficacy [31] but otherwise remains absent from HIV prevention research.

1.3.2 Studies of Atherosclerosis and Coronary Heart Disease

The competing risks framework we have developed to motivate use of the compound Poisson distribution for modeling frailty is also applicable to research on the relationship between atherosclerotic lesions and risk of coronary heart disease (CHD) related events. Each lesion presents a distinct exposure process that may result, eventually, in a CHD event. Then, assuming the risks posed by each lesion are independent of each other, an individual's risk of a CHD event is the sum of risks across all lesions. In comparison to our previous application, however, in this setting, it is possible to observe the number of exposure processes. Computed topography (CT) can be used to identify calcified lesions as well as lesion characteristics such as location, volume, diameter, and the amount of coronary artery calcium present. These lesion-level factors likely influence the risk associated with each lesion. Individual-level factors such as hypertension, diabetes, or smoking habits also affect an individual's risk of an event.

An ideal model for the risk of CHD-related events would reflect both individual- and lesion-level sources of risk. Yet, it is not possible to estimate the effect of lesion-level factors on CHD event-free survival with existing frailty methods because events are not observed at the lesion-level but, instead, are observed at the level of the individual and the lesion that experienced the event is unknown. As a consequence, survival must be modeled on the individual-level, and the individual-level model is not able to accommodate lesion-level covariates. So, lesion-level characteristics are either ignored or some aggregate measure of these characteristics are used. As an alternative, we outline ideas for a hierarchical frailty model that models the risk associated with each lesion and sums over that risk to obtain an individual's frailty.

1.4 Notation

We briefly detail the conventions followed for notation in this dissertation. Observed data are typically denoted with lower case Latin letters. For example, time-to-event is denoted by t. The only exception to this rule is δ, which is commonly used to indicate status in time-to-event data, so we adopt that convention here as well. Latent variables are represented

by uppercase Latin letters. As seen above, Z denotes unobserved frailty.

All model parameters are represented by lowercase Greek letters. For example, β is used throughout to represent the log-hazard ratio and λ(t) denotes the baseline hazard function. Note, the cumulative baseline hazard function is denoted by $\Lambda(t)$, as is convention.

Prior distributions for model parameters are indicated with $\boldsymbol{\pi}(\cdot)$ and hyperparameters are represented with uppercase Greek letters (many of which appear identical to uppercase Latin letters). In Chapter 2, we use hyperpriors to correlate baseline hazard functions across strata. In this case, uppercase Greek letters are only used at the highest level of prior specification.

Finally, Latin letters are also used to index data as well as model parameters and their associated hyperparameters. For example, the vector of covariates in a Cox regression, $\mathbf{x}$, is indexed by p: $\mathbf{x} = [x_1, \ldots, x_P]$. The corresponding log hazard ratio vector, $\boldsymbol{\beta}$, also has P components: $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_P]$. It follows that the hyperparameters for the priors on $\boldsymbol{\beta}$, $\boldsymbol{\pi}(\boldsymbol{\beta}) = Normal(\mathbf{M}, \boldsymbol{\Sigma}^2)$, have P elements as well: $\boldsymbol{\pi}(\beta_1) = Normal(M_1, \Sigma_1^2), \ldots, \boldsymbol{\pi}(\beta_P) = Normal(M_P, \Sigma_P^2)$, assuming independent normal priors for each component of $\boldsymbol{\beta}$.

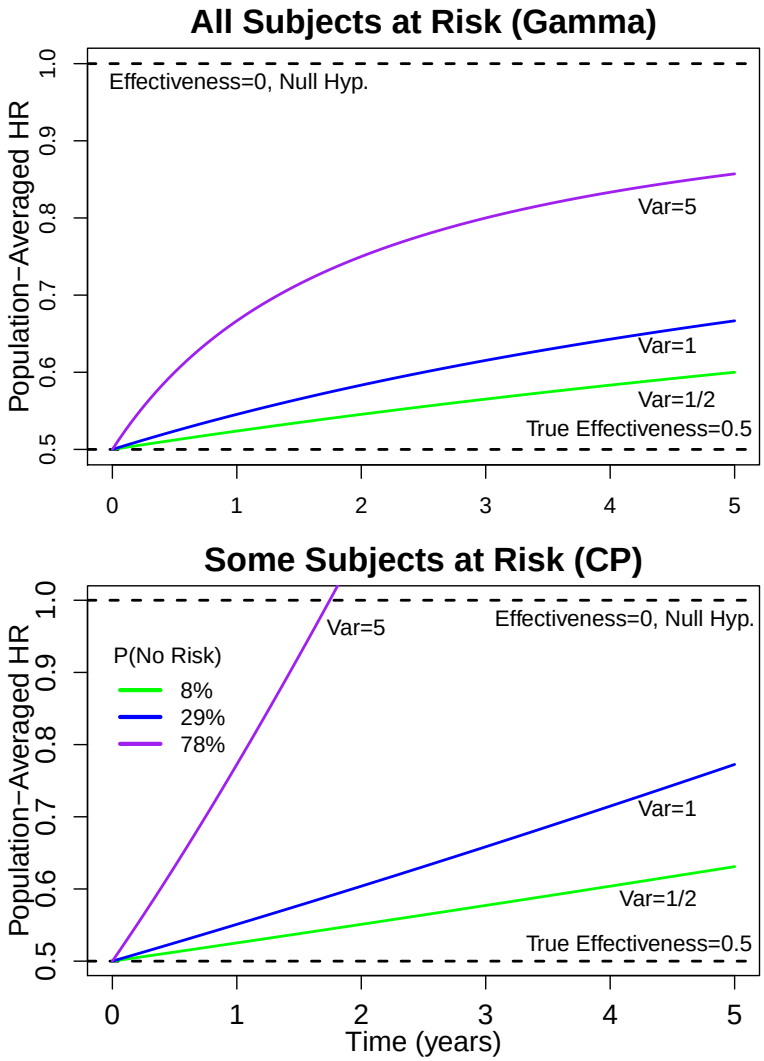

Figure 1.1: Population-averaged hazard ratio when heterogeneity follows a gamma (upper panel) or compound Poisson (CP) (lower panel) distribution.

Chapter 2

BAYESIAN HIERARCHICAL COMPOUND POISSON FRAILTY MODEL

2.1 Introduction

Modeling heterogeneity in risk is necessary in order to estimate the individual-level effect of an intervention, exposure, or condition on survival. Including a frailty term in a Cox model is an approach that accounts for heterogeneity while maintaining proportional hazards. The gamma distribution is a common choice for the frailty distribution, primarily due to mathematical convenience rather than biological rationale. As an alternative, we prefer to select a frailty distribution that better reflects the sources of heterogeneity in risk. In this chapter, we focus on the use of the compound Poisson distribution. We contribute to existing literature by offering an alternative, application-motivated, and hierarchical definition of the distribution. This research is also the first to demonstrate Bayesian estimation of a compound Poisson frailty model.

Using the compound Poisson distribution to model frailty was first suggested by Aalen (1988), although the general effect of heterogeneity on survival estimates was recognized previously [25, 32, 51]. This recommendation for a compound Poisson frailty model was motivated by the fact that the divergence of population-level estimates from individual-level effects that occur the presence of heterogeneity is more drastic when some portion of the population is not at risk of an event. (Aalen's presentation is replicated in Chapter 1, Figure 1.1.) Modeling frailty with the compound Poisson distribution both allows for a proportion of the population to have no risk as well as for variability in the level of risk among those at risk. Similar models, such as cure-rate [10] and mover-stayer [51] models, accommodate a cured or surviving fraction but do not provide a unified model for heterogeneity among all participants that guarantees proportional hazards.

The definition of a compound Poisson random variable, the sum of a Poisson-distributed

random number of Gamma-distributed random variables, invites explicit biological interpretation. Despite Aalen's early advocacy for the compound Poisson distribution's use in frailty modeling, Aalen did not emphasize the possible interpretations for the distribution, stating in a 1992 paper that "one should obviously not interpret it too literally in any given practical case" (p 953). Instead of a more natural parameterization of the model, one that reflected the distribution's construction and enabled straightforward estimation of the underlying Poisson and gamma distributions, Aalen's initial estimation of the model relied on the Laplace transformation of the compound Poisson density, a construction in which the parameters have no meaningful interpretation. Later applications of the compound Poisson frailty model have followed this example [38, 39, 43] while applications that highlight the model's interpretation are rare [9].

We argue that a more explicit interpretation of the frailty model is an advantage of using the compound Poisson distribution, which reflects a risk mechanism common in medical contexts where an individual's risk of an event is the result of independent, competing exposure processes. Examples include patterns of risk behavior resulting in HIV infection, atherosclerotic lesions causing a cardiovascular event, and tumors leading to cancer recurrence. The number of exposure processes for each individual, frequently either measured with error or unknown, is modeled as a Poisson random variable, which allows for a participant to have no risk (i.e. no exposure). Each exposure process is assumed to present a constant level of risk over time, which is modeled as a gamma random variable. Separate exposure processes pose independent, competing risks; given exposure, one of the processes eventually leads to an event. It follows that an individual's risk of an event during a specified time period is the sum of these competing risks.

As an example, consider the risk of HIV seroconversion. First, an individual must have sexual contact with an HIV positive partner to have any risk of seroconversion. This contact occurs in the context of that individual's patterns of sexual behavior, which are not fully observed. Second, the risk of seroconversion is not equal for each exposure or sexual partnership, rather, it depends on the number and type of acts, viral load of the infected partner(s), and other factors that may be unobserved. Here, we can consider separate behavior patterns to be distinct exposure processes that present constant risk over time. For

example, unprotected vaginal sex with an infected partner constitutes one exposure process, while unprotected anal sex with the same partner may be another. Of course, the number of exposure processes for any individual is unknown, but is assumed to follow a Poisson distribution. For individuals with nonzero exposure processes, the risk associated with each exposure process is a random variable distributed continuously over the positive half-line. Finally, we assume that the risks associated with each exposure process are independent competing risks, that is, we assume that the exposure processes act independently to cause seroconversion, such that the sum of these competing risks represents an individual's risk of seroconversion during a specified interval of time.

The chapter proceeds as follows. First, we propose a hierarchical definition of the compound Poisson frailty model and outline Bayesian model estimation procedures, including consideration for prior distributions. Model assessment is also briefly discussed. Then, model performance is evaluated via simulation. Finally, we illustrate the use of this model to estimate the individual-level effectiveness of an intervention in an HIV prevention trial.

2.2 Methods

We first present a survival model with frailty before defining the compound Poisson distribution. Our survival model follows the typical formulation for a frailty model where a frailty term, Z_i, scales the hazard for individual i, $i = 1, \ldots, n$. We define the ith participant's hazard function as $\lambda_i(t|\mathbf{x}_i) = \lambda(t)\, Z_i \exp(\mathbf{x}_i'\boldsymbol{\beta})$ where $\lambda(t)$ is a baseline hazard function common to all participants and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)$ is a P-length vector of the log-hazard ratio associated with covariates $x_1, \ldots, x_P$, respectively. The vector $\mathbf{x}_i = (x_{1i}, \ldots, x_{Pi})$ is the observed values of these covariates for participant i. It follows that the subject-specific density (2.1) and survival function (2.2) are:

$$f(t|\mathbf{x}_i, Z_i) = \lambda(t)\, Z_i \exp(\mathbf{x}_i'\boldsymbol{\beta})\, S(t|x_i, Z_i) \tag{2.1}$$

$$S(t|\mathbf{x}_i, Z_i) = \exp\Big(- Z_i \exp(\mathbf{x}_i'\boldsymbol{\beta})\, \Lambda(t)\Big), \tag{2.2}$$

where $\Lambda(t) = \int_0^t \lambda(t)\, dt$ is the cumulative baseline hazard at time t.

The distribution of frailty in a survival model is frequently chosen based on mathematical convenience. Instead, we propose the use of the compound Poisson distribution, which

reflects the sources of heterogeneity in risk in contexts where an individual's risk of an event is the result of independent, competing exposure processes. The number of exposure processes, N_i, for the ith participant, $i = 1, \ldots, n$, follows a Poisson distribution:

$$N_i \sim Poisson(\rho).$$

Each exposure process presents a constant, unobserved risk over time and the risk arising from each exposure process varies, such that the ith participant's mth exposure process is associated with a hazard for seroconversion, $\lambda(t)U_{im}$, where

$$U_{im} \sim Gamma(\eta, \nu) \qquad m = 1, \ldots, N_i$$

and $Gamma(\eta, \nu)$ is a gamma distribution with shape parameter η and rate parameter ν. Multiple exposure processes for an individual pose competing risks. Assuming these risks are independent, that is, that each process acts independently to cause an event, the risk of seroconversion is additive across processes and a participant's baseline risk of seroconversion at time t equals $\lambda(t) \sum_{m=1}^{N_i} U_{im}$. We define Z_i such that

$$Z_i \quad = \quad \begin{cases} \sum_{m=1}^{N_i} U_{im} \sim \text{Gamma}(N_i\eta, \nu) & \text{if } N_i > 0 \\ 0 & \text{if } N_i = 0. \end{cases}$$

Formally, the frailty term Z_i follows a compound Poisson distribution [1]. To ensure identifiability of the frailty model, we set the expectation of Z_i, $E(E(Z_i|N_i)) = \rho\eta/\nu$, equal to one by defining ν in terms of the other frailty distribution parameters, ρ and η: $\nu = \rho\eta$. We also note that the probability that the frailty for an individual is greater than zero, $P(Z_i > 0)$, is equal to $P(N_i > 0) = 1 - \exp(-\rho)$. In application, this probability corresponds to the proportion of participants with some positive risk of an event during the time frame under consideration.

Having defined an appropriate survival model and frailty distribution, we next define the model likelihood given time-to-event data. For each participant, we observe (t_i, δ_i) where t_i is potentially censored time to event and δ_i is an indicator variable equaling one if an event

is observed and zero otherwise. The likelihood for the proposed model is:

$$
\begin{aligned}
&L\big(\boldsymbol{\beta}, \lambda(t), \rho, \eta | (t_i, \delta_i, \mathbf{x}_i, N_i, Z_i), i = 1, \ldots, n\big) \\
&\quad = \prod_{i=1}^{n} f(N_i|\rho)\, f(Z_i|N_i, \rho, \eta)\, f(t_i, \delta_i|\mathbf{x}_i, Z_i, \boldsymbol{\beta}, \lambda(t)) \\
&\quad = \prod_{i=1}^{n} f(N_i|\rho_i) \left[f(Z_i|N_i, \rho, \eta)\, f(t_i, \delta_i|\mathbf{x}_i, Z_i, \boldsymbol{\beta}, \lambda(t)) \right]^{\mathbf{1}_{[N_i>0]}} \\
&\qquad \times \left[f(Z_i|N_i, \rho, \eta)\, f(t_i, \delta_i|\mathbf{x}_i, Z_i, \boldsymbol{\beta}, \lambda) \right]^{\mathbf{1}_{[N_i=0]}} \\
&\quad = \prod_{i=1}^{n} \frac{\rho^{N_i}}{\Gamma(N_i+1)} \exp(-\rho) \Big[\frac{\rho \eta^{N_i \eta}}{\Gamma(N_i \eta)} Z_i^{N_i \eta - 1} \exp(-Z_i \rho \eta) \\
&\qquad\qquad \big(\lambda(t) Z_i \exp(\mathbf{x}_i' \boldsymbol{\beta})\big)^{\delta_i} \exp\big(- Z_i \exp(\mathbf{x}_i' \boldsymbol{\beta})\, \Lambda(t_i)\big) \Big]^{\mathbf{1}_{[N_i>0]}} \\
&\qquad \times \left[\boldsymbol{\Delta}_0(Z_i)\, \boldsymbol{\Delta}_\infty(t_i) \right]^{\mathbf{1}_{[N_i=0]}}, \qquad (2.3)
\end{aligned}
$$

where $\boldsymbol{\Delta}_0(Z_i)$ and $\boldsymbol{\Delta}_\infty(t_i)$ indicate point mass for frailty, Z_i, at zero and time-to-event, t_i, at infinity for participant i with $N_i = 0$, i.e. no risk.

Before discussing prior elicitation and estimation procedures for this model, we make a few remarks regarding interpretation, model specification, and stratification. First, we note that the interpretation of the hazard ratio, HR, $\exp(\beta_p)$ for $p = 1, \ldots, P$, differs between the Cox model and compound Poisson frailty model. In the Cox model, the HR quantifies the change in risk for a population given a one unit increase in covariate x_p, but inclusion of a compound-Poison distributed frailty term changes this interpretation. Instead, in the compound Poisson frailty model, the HR refers to the change in risk for an at-risk individual. Participants with no risk do not contribute to the estimate. For example, in an HIV prevention trial to assess the effectiveness of an intervention, the Cox model HR compares the risk of seroconversion for a population that receives the intervention to one that does not while the proposed model's HR measures the effectiveness of the intervention for an individual with some positive risk of seroconversion.

In contrast to interpretation of the compound Poisson frailty model's HR, interpretation of the baseline hazard function in the compound Poisson model applies to all individuals. Given that the hazard at time t for individual i is $Z_i \lambda(t)$ and that the expectation of Z_i is one, the baseline hazard, $\lambda(t)$, represents the average baseline hazard at time t among

all participants, including those with no risk (for whom $Z_i = 0$). It is possible, however, to calculate the baseline hazard among only those individuals with non-zero risk of an event. The average baseline hazard among participants at risk is a function of $\lambda(t)$ and the proportion of population at risk: $E(Z_i\lambda(t) \mid Z_i > 0) = \lambda(t)/(1 - \exp(-\rho))$.

The form of the baseline hazard in this model's presentation is intentionally general and can accommodate a variety of specifications, several of which are discussed here. Specifying a constant baseline hazard function, for example, assumes that each participant has a constant risk of an event throughout the trial with the level of risk for individuals varying according to a compound Poisson distribution. Several previous applications of the compound Poisson frailty distribution have modeled the baseline hazard with a Weibull function, which assumes a monotonic trend in hazard for an individual [1, 2, 38, 39]. If we do not want to make assumptions about the form of the baseline hazard, less parametric alternatives include a piecewise constant hazard function, spline representation of the baseline hazard, or a Gamma process model [27]. Yet, estimation of semi- and non-parameteric hazard functions is difficult without a large number of events observed. When considering these and other options, the form of the baseline hazard function should be chosen based on scientific rationale and supported by examination of the data. We note that, in a Cox model without frailty, model estimation with a partial likelihood approach does not require a baseline hazard function to be specified. Inclusion of frailty, however, precludes this option.

For some applications, we would consider stratifying the model by site, as is typically done for analysis of data from multi-site clinical trials. Site-stratification would involve estimating the baseline hazard, $\lambda(t)$, as well as the parameters of the frailty distribution, ρ and ν, separately for each study site. Consequently, the proportion at risk would vary across site, as would the distribution of frailty among those at risk. The average frailty within each site would also be constrained to equal one. Finally, the baseline hazard would also vary across sites, which allows the average baseline hazard to be higher in some sites and lower in others.

2.2.1 Prior Elicitation

To complete Bayesian specification of the proposed model, we select prior distributions for model parameters.

There are several reasonable priors for ρ, the mean parameter for the number of exposure processes. If prior knowledge about the number of relevant exposure processes in the study population were available, that information could be incorporated into a gamma prior for ρ: $\boldsymbol{\pi}(\rho) = Gamma(O_\rho, T_\rho)$ where $Gamma(O, T)$ is the Gamma distribution with shape O and rate T. Alternatively, the prior on ρ could be specified to reflect prior knowledge about the proportion at risk in the study population. Recall that the proportion of participants at risk, which we denote $P(Risk)$, is also a function of ρ: $P(Risk) = P(N_i > 0) = 1 - \exp(-\rho)$. We specify $P(Risk) \sim Beta(A, B)$, where $Beta(A, B)$ is the beta distribution with shape parameters A and B. The prior distribution for ρ is then:

$$\boldsymbol{\pi}(\rho) = \frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)} \left(1 - \exp(-\rho)\right)^{A-1} \exp(-\rho)^B. \tag{2.4}$$

Setting $A = 1$ and $B = 1$ is equivalent to a Uniform(0,1) distribution on the proportion at risk, $P(Risk)$, and an exponential prior with mean one on the expected number of exposure processes, ρ. This prior is appropriate when we do not have, or do not want to specify, any information about the proportion at risk or the number of risk processes for participants. Alternatively, a more informative prior for ρ could be determined using previous research on the study population or even preliminary trial data. We discuss this approach in Section 2.4.1.

The gamma distribution is a natural choice for the prior distribution on η, which determines the level of risk associated with each exposure process. Data from previous studies could also be used to inform the prior since the number and timing of events observed would be more likely under certain values of η.

Prior selection for the baseline hazard function will depend on the form chosen for $\lambda(t)$. For example, under the assumption of a constant baseline hazard, $\lambda(t) = \lambda$ for all t, the gamma distribution is a conjugate prior, and we set $\boldsymbol{\pi}(\lambda) = Gamma(O_\lambda, T_\lambda)$. Similarly, when the baseline hazard is assumed to be piecewise constant, we place independent gamma priors on the constant hazard for each of S intervals: $\boldsymbol{\pi}(\lambda_s) = Gamma(O_{\lambda_s}, T_{\lambda_s})$, $s =$

$1, \ldots, S$. A gamma process prior is another appropriate choice for a piecewise constant baseline hazard. It is also possible to specify a prior on the baseline hazard that is correlated with the prior on the proportion at risk. We present details about a correlated prior selected for our particular application in Section 2.4.1.

Finally, we choose a normal prior for each component of $\boldsymbol{\beta}$, the P-length vector of log-hazards ratios, because there is no conjugate prior available, a normal distribution covers the possible domain for the parameter, and prior knowledge can be easily incorporated: $\boldsymbol{\pi}(\beta_p) = N(M_{\beta_p}, \Sigma^2_{\beta_p})$, $p = 1, \ldots, P$, where $N(M, \Sigma^2)$ is a normal distribution with mean M and variance Σ^2. Setting the mean at zero and the variance to be relatively large for a particular component of $\boldsymbol{\beta}$ would provide an uninformative prior, which is typically appropriate in a clinical trial setting. Alternatively, previous research about a covariate's effect on risk could be used to determine a more informative prior.

2.2.2 Posterior Estimation Procedure

After specifying priors for all model parameters, we can define the joint posterior density of the parameters and latent variables as the product of the model likelihood in Equation (2.3) and each parameter's prior density:

$$\begin{aligned}\mathbf{p}\Big(\boldsymbol{\beta}, \lambda(t), \nu, \rho, \mathbf{N}, \mathbf{Z} \big| \big((t_i, \delta_i, \mathbf{x}_i), i = 1, \ldots, n\big),\, \boldsymbol{\Theta}\Big) \\ \propto L(\boldsymbol{\beta}, \lambda(t), \nu, \rho | (t_i, \delta_i, \mathbf{x}_i, N_i, Z_i), i = 1, \ldots, n) \times \boldsymbol{\pi}(\beta, \lambda(t), \rho, \eta | \boldsymbol{\Theta}). \qquad (2.5)\end{aligned}$$

where $\mathbf{p}(\cdot|\text{data})$ denotes the joint posterior density, $\mathbf{N}$ and $\mathbf{Z}$ are the vectors of latent variables $[N_1, \ldots, N_n]$ and $[Z_1, \ldots, Z_n]$, and $\boldsymbol{\pi}(\cdot|\boldsymbol{\Theta})$ denotes the joint prior density for model parameters conditional on hyperparameters $\boldsymbol{\Theta}$.

We use a Gibbs sampling algorithm [21] to sample from each parameter's full conditional posterior distribution, all of which follow from the joint posterior and are detailed in Appendix B. Posterior samples for each parameter are updated in turn; slice sampling [40] is used when conjugate priors are not available.

At each iteration of the sampling algorithm, latent variables N_i and Z_i, the number of exposure processes and frailty for the ith participant, are also sampled for every $i = 1, \ldots, n$. Conditional likelihoods for N_i and Z_i are obtained from the model likelihood in Equation

(2.3). In order to sample N_i, it is necessary to remove Z_i from the conditional likelihood of N_i, as an individual's frailty is dependent on her number of exposure processes, per our hierarchical model. So, we sample N_i from its distribution conditional on observed data and all parameters but marginalized over Z_i:

$$\begin{aligned} f(N_i|\rho,\, \eta,\, \lambda(t),\, \boldsymbol{\beta}, t_i, \delta_i, \mathbf{x}_i) \\ &\propto \frac{\rho^{N_i}}{\Gamma(N_i+1)} \int_0^\infty \Big(\frac{\rho\eta^{N_i\eta}}{\Gamma(N_i\eta)} Z_i^{N_i\eta-1} \exp(-Z_i\rho\eta) \\ &\qquad \big(\lambda(t) Z_i \exp(\mathbf{x}_i'\boldsymbol{\beta})\big)^{\delta_i} \exp\big(- Z_i \exp(\mathbf{x}_i'\boldsymbol{\beta})\Lambda(t_i)\big)\Big)\, dZ_i \\ &\propto \frac{\rho^{N_i}}{\Gamma(N_i+1)} \Big(\frac{\rho\eta}{\rho\eta + \exp(\mathbf{x}_i'\boldsymbol{\beta})\Lambda(t_i)} \Big)^{N_i\eta} \Big(\frac{N_i\eta}{\rho\eta + \exp(\mathbf{x}_i'\boldsymbol{\beta})\Lambda(t_i)} \Big)^{\delta_i} \end{aligned} \tag{2.6}$$

The approach for sampling N_i from Equation (2.6) depends on whether seroconversion was observed. When $\delta_i = 0$, i.e. seroconversion was not observed for participant i, the conditional distribution of N_i marginalized over Z_i is proportional to a Poisson distribution with mean $\rho\big(\rho\eta/(\rho\eta + \exp(\mathbf{x}_i'\boldsymbol{\beta})\Lambda(t_i))\,\big)^\eta$. We draw from this Poisson distribution to obtain N_i. When $\delta_i = 1$, i.e. seroconversion was observed, we use a Metropolis-Hastings type algorithm to sample N_i from Equation (2.6), with the constraint that N_i must be positive. (Since seroconversion occurred, the participant must have had at least one exposure process.)

After sampling the number of exposure processes for an individual, we sample frailty. If the sampled number of exposure processes for participant i, N_i, is nonzero, the conditional likelihood of frailty, Z_i, is proportional to a gamma distribution with shape $= N_i\eta + \delta_i$ and rate $= \rho\eta + \exp(\mathbf{x}_i'\boldsymbol{\beta})\Lambda(t_i)$. Hence, to obtain Z_i conditional on N_i, observed data, and all parameters, we simply draw from this gamma distribution. If the sampled number of exposure processes is zero, then the frailty is necessarily zero.

2.2.3 Model Assessment

Predictive accuracy of Bayesian hierarchical models is commonly assessed with the deviance information criterion (DIC) [46]. However, some studies have found that the DIC does not provide correct comparison for high dimension latent variable models (i.e. one or more latent parameters per observation) by demonstrating that DIC calculations using the like-

lihood conditional on posterior samples of latent variables does not lead to correct model comparison [16, 37]. Integrating over latent variables when calculating DIC may improve performance, but to do so is often difficult and computationally expensive. In our simulations, we found that DIC gave inconsistent results for the proposed model, as did the Watanabe-Akaike information criteria, a common alternative to the DIC that relies on the same conditional likelihood [52]. Finally, while criteria based on the posterior predictive distribution, such as predictive concordance and the L-measure, offer an alternative approach, these measurements can not be obtained for survival data with a cured fraction [20, 27]. As a consequence, there is not a common measure of predictive accuracy appropriate for our context.

In this chapter and throughout this dissertation, we will assess model fit using Bayesian p-values for survival data from the posterior predictive distribution [22]. Specifically, at each iteration of the Gibbs sampler, frailty and event times are generated using current posterior samples of all model parameters. The probability of survival between, for example, 6 and 12 months is then compared to observed survival in that same time period. The Bayesian p-value for an interval is the proportion of posterior replicates where generated survival was below observed survival. The optimal Bayesian p-value is around 0.5, indicating that half of the predicted survival is above what was observed and half is below. A Bayesian p-value below 0.5 indicates that survival is over-estimated by the model, and one above 0.5 indicates that survival is under-estimated. Unfortunately, there is no universally recognized threshold at which a Bayesian p-value is considered too far from 0.5, limiting the utility of this measure. Limitations of Bayesian p-values are discussed in Section 4.2.

2.3 Simulations

2.3.1 Methods

We performed a series of simulations to assess the properties of the proposed model. For each simulation, data were generated from the proposed model for a sample size of $n = 3,000$, divided evenly between intervention and control groups. Compound Poisson frailties were generated using a shared value of ρ, the mean parameter for the number of exposure

processes. Separate studies were done for a range of values for ρ, which corresponded to varying proportions at risk of an event, $1 - \exp(-\rho) = 0.15, 0.25, \ldots, 0.55$. The shape parameter for the frailty distribution, η, was fixed at 2 for all simulation studies. (We chose to focus on the effects of varying ρ in these simulations instead of η because ρ has a more direct and interesting interpretation.) Event times were generated based on the simulated compound Poisson-distributed frailties, a constant baseline hazard function of 10 events per 100 person-years (PY), $\lambda(t) = 0.10$, and an intervention effectiveness of 50%, $\exp(\beta) = 0.5$. (Values of η and λ were selected in order to observe events at a pace comparable to ARV-based HIV prevention trials.) For each simulated data set, censoring was done after a pre-specified number of events- 90, 180, 270, and 360, that is, event rates of 3%, 6%, 9%, and 12%- so that differences in effectiveness estimates across studies would not be driven by the number of events observed. Event rates as low as 3% are uncommon and typically only seen when a trial is stopped early, as in [8]. Event rates as high as 12% are also uncommon but possible (see [4]). An event rate of about 6% was observed in the trial examined in the application later in this chapter.

In this set of simulations, the prior mean was set equal to the true value for all parameters, except β, yet prior distributions were broad enough to allow for a wide range of values for posterior samples. The prior on β was non-informative. We set the following priors on model parameters.

$$\text{Prior:} \begin{cases} \beta \sim N(0, 100) \\ \lambda \sim Gamma(0.05 \times 50, 50) \\ P(Risk) = 1 - \exp(-\rho) \sim Beta(2 \times \frac{P(Risk)}{1-P(Risk)}, 2) \\ \eta \sim Gamma(1, \frac{1}{2}) \end{cases} \tag{2.7}$$

At each replication, posterior medians for model parameters and surrounding 95% highest posterior density (HPD) intervals were identified for the proposed compound Poisson frailty model. Calculation of HPD intervals was performed with the `HPDinterval` function in the `coda` package available for `R`. Cox model estimates of intervention effectiveness and model-based 95% confidence intervals were estimated using the `coxph` function in the `survival` package available in `R`. Simulations for each scenario were replicated 250 times.

Posterior samples were the results of 25,000 replications of the Gibbs sampling algorithm, of which 10,000 were discarded for burn-in. Convergence was assessed for a few replications of each simulation by examining trace plots and cumulative quantile plots.

Model performance under various prior specifications was also examined. We repeated the simulation for 25% of participants at risk and an event rate of 6% under three prior specification scenarios. (This event rate and proportion at risk were chosen because they were close to the event rate and seroprevalence, respectively, for the application in Section 2.4.) The scenarios examined were as follows:

(I) Incorrect prior on proportion at risk: The prior distribution for the proportion at risk had a mean of 50%, $\boldsymbol{\pi}(1 - \exp(-\rho)) = Beta(2, 2)$, twice the true value of 25%.

(II) Incorrect prior on baseline hazard: The prior distribution on the constant baseline hazard was centered at 10 events per 100 person-years, $\boldsymbol{\pi}(\lambda) = Gamma(0.1 \times 50, 50)$, twice the true value of 5 events per 100 person-years.

(II) Non-informative priors: Uniform priors were placed on all parameters, including a flat prior on the proportion at risk, $\boldsymbol{\pi}(1 - \exp(-\rho)) = Beta(1, 1)$, instead of on the mean number of exposure processes, ρ.

For these simulation scenarios, we identified posterior medians for all model parameters as well as coverage of quantile-based 95% credible intervals. We also calculated Bayesian p-values for survival at six month intervals using the method described in Section 2.2.3. For each simulation replication, event times generated at each iteration of the Gibbs sampler were compared to those observe for that replication. 500 replications were performed for each simulation study.

2.3.2 Results

Figure 2.1 shows effectiveness (1-HR) estimates from the proposed model using the priors in Equation (2.7) and the Cox model given the proportion of participants at risk (upper panel) as well as the coverage of intervals around the estimated effectiveness (lower panel)

	HR		P(Risk)		η		λ	
P(Risk)	Est	Cov	Est	Cov	Est	Cov	Est	Cov
0.15	0.49	95%	0.17	100%	1.8	100%	5.1	95%
0.25	0.51	96%	0.28	98%	1.8	100%	5.1	97%
0.35	0.50	96%	0.38	96%	1.8	100%	5.0	97%
0.45	0.50	96%	0.48	96%	1.8	100%	5.0	96%
0.55	0.50	96%	0.60	90%	1.8	100%	4.9	98%
Truth			0.5		5 / 100PY		2	

Table 2.1: Average posterior median for model parameters and coverage of 95% HPD intervals from 250 replications at 6% event rate.

at a range of event rates and proportion at risk. We see that the proposed estimation procedure produced unbiased estimates and 95% HPD intervals with nominal coverage for β. As expected, effectiveness estimates from the Cox model underestimated individual-level effectiveness and had less than nominal coverage. Estimate attenuation and inadequate coverage were more pronounced as the proportion of the sample at risk declined. When the proportion at risk was higher and the number of events lower, however, estimates from the Cox and compound Poisson models were similar.

Table 2.1 gives the average posterior median and coverage of the 95% HPD intervals for the compound Poisson frailty model when data were simulated with a 6% event rate. We see that estimates for the baseline hazard, in addition to those for the HR, were unbiased. 95% HPD intervals for model parameters were generally conservative, with the possible exception of β. We also see that the proposed model slightly overestimates the proportion at risk, $1-\exp(-\rho)$ while underestimating η. It is likely that η was underestimated because the prior for that parameter, while centered at the true value, was quite skewed in order to allow for more variability in the posterior. Furthermore, as discussed below, the posterior distribution for η follows the prior closely. Since η was underestimated, it is not surprising that the proportion at risk was overestimated.

Scenario	HR		$P(Risk)$		η		λ	
	Est	Cov	Est	Cov	Est	Cov	Est	Cov
(0): Correct priors	0.50	96%	0.28	99%	1.8	100%	5.0	97%
(I): Incorrect $\boldsymbol{\pi}\big(P(Risk)\big)$	0.52	96%	0.45	76%	1.3	99%	4.7	94%
(II): Incorrect $\boldsymbol{\pi}(\lambda)$	0.52	96%	0.27	99%	1.7	100%	5.3	96%
(III): Non-informative priors	0.52	94%	0.32	45%	61	17%	4.6	78%
Truth	0.5		0.25		2		5 / 100PY	

Table 2.2: Average posterior median for model parameters and coverage of 95% credible intervals under various prior specification scenarios from 500 replications at 6% event rate.

Scenario	Bayesian p-values		
	6 mo.	12 mo.	18 mo
(0): Correct priors	0.53	0.48	0.48
(I): Incorrect $\boldsymbol{\pi}\big(P(Risk)\big)$	0.46	0.47	0.52
(II):Incorrect $\boldsymbol{\pi}(\lambda)$	0.60	0.51	0.49
(III): Non-informative priors	0.43	0.48	0.55

Table 2.3: Bayesian p-values for survival at 6 month intervals under various prior specification scenarios from 500 replications at 6% event rate.

The prior and posterior distributions for model parameters from a single replication of the simulation with 25% at risk and 6% event rate are shown in Figure 2.2. We see that posteriors for the log-HR (β), baseline hazard (λ) and proportion at risk ($1-\exp(-\rho)$) differ considerably from their prior distributions. The posterior for η, however, is similar to its prior distribution, suggesting that this parameter may be more sensitive to prior selection.

Simulation results under non-informative and misspecified priors are given in Tables 2.2 and 2.3. We see that hazard ratio estimation is quite robust to prior misspecification. While effectiveness may be slightly underestimated, coverage of 95% credible intervals is

still nominal. Bayesian p-values for survival at 6, 12, and 18 months are also close to the optimal value of 0.50, indicating that survival data generated under the joint posterior is consistent with observed survival.

Posterior estimates of other model parameters are less robust. In misspecification scenario (I), the prior distribution on the proportion at risk is centered at 50%, twice as high as the true proportion at risk of 25%. As a result, the proportion at risk was overestimated. Subsequently, the baseline hazard, λ, and the shape parameter for the risk associated with each exposure process, η, are both underestimated. When the expectation of prior distribution on the baseline hazard is twice as high as the true value, as in scenario (II), the baseline hazard is over-estimated while estimation of ρ and η appears unaffected. When non-informative priors are used, as in scenario (III), η is greatly overestimated. Since the mean frailty is constrained to one and the rate parameter for the risk associated with each exposure process is the product of ρ and η, drastic increases in η can be balanced by decreases in ρ and an increase in their product corresponds to a more narrow distribution for positive frailties. In this scenario, while the average posterior mean for the proportion at risk is not far from the true value (32% vs. 25%), variability in the posterior distribution for the proportion at risk across simulations is so great that less than half of the 95% credible intervals cover the true value. Placing some prior constraints on η would likely reduce the variability in ρ, even if the prior on the proportion at risk remained non-informative. Given the similarity between the prior and posterior distributions seen above, it may help to restrict the range of η while specifying a relatively flat prior within that range if little information is available to determine an appropriately informative one. Despite biased estimates of ρ, η, and λ and poor coverage of their surrounding intervals, estimation of the HR is only slightly biased and survival data generated under the posterior is similar to that which was observed.

We note that, although identifiability of this class of frailty models is well established [19, 51], the properties of estimates under this particular parameterization of the compound Poisson distribution have not yet, to our knowledge, been demonstrated theoretically. Our simulation-based exploration of estimate properties suggests that identifiability of frailty distribution parameters may be difficult in the absence of constraints or use of informative

priors.

2.4 Application

2.4.1 Data and Analysis

We applied the model to data from HPTN 035, a completed Phase II/IIb trial to evaluate the safety and effectiveness of BufferGel and 0.5% PRO 2000 microbicide gels in preventing male-to-female transmission of HIV [5]. Beginning in February 2005, 3,099 women were enrolled in the study from seven sites in Sub-Saharan Africa and one site in Philadelphia, USA. Participants were randomized to one of four arms: BufferGel, 0.5 % PRO 2000, placebo gel, and no gel and tested quarterly for seroconversion for twelve to thirty months. The primary analysis used the Cox proportional hazards model to estimate the hazard ratio of seroconversion between each candidate gel and both the placebo gel and no gel. BufferGel did not reduce the risk of seroconversion, nor did placebo gel in comparison to no gel. The 0.5% PRO 2000 gel showed a 28% reduction in HIV acquisition compared to the placebo gel, but this result was not statistically significant at the 0.05 level.

In our analysis, we estimated the effectiveness of 0.5% PRO 2000 compared to placebo gel using data from the seven African sites. Our analysis focused on 0.5% PRO 2000 gel because it showed some effect in the primary analysis. Data from Philadelphia was excluded due to too few events for a stratified analysis. Following the primary analysis from the trial, we excluded participants who were infected with HIV at baseline but did not seroconvert until after enrollment. For participants without seroconversion observed, we used the last negative test as the censored event time. For participants with seroconversion observed, we sampled event time at each iteration of the Gibbs sampler using slice sampling techniques and a uniform prior over the interval covering the time between a participant's last negative and first positive serology test. Cox model estimation was performed for comparison, and we used the midpoint between last negative and first positive test for event time of seroconverters in that analysis. (By comparison, Abdool Karim et al (2011) used the time of first positive test was used as the event time in the primary analysis.)

We estimated the individual-level effectiveness of 0.5% PRO 2000 gel in reducing HIV se-

roconversion risk relative to placebo gel under two versions of the proposed compound Poisson model: non-stratified and stratified by site. For the non-stratifed model, we estimated a single, shared value for the baseline hazard, $\lambda(t)$, and parameters of the frailty distribution, ρ and η. In the stratified model, these parameters were estimated separately for each site. Estimates of the intervention's effect for an exposed individual from the non-stratified and stratified compound Poisson frailty models were compared to population-averaged estimates of effectiveness from the non-stratified and stratified Cox models, respectively. 95% highest posterior density credible intervals were identified for the compound Poisson frailty models while model-based 95% confidence intervals were calculated for the Cox models.

For both variations of the compound Poisson frailty model, we used screening data to set the prior on the mean number of exposure processes, ρ, via assumptions about the proportion at risk, as described in Section 2.2.1. We set the expectation of a beta prior for the proportion of the study population at risk equal to the proportion of individuals screened for the study who tested seropositive using the following rationale: The screening prevalence in the eligible population may provide a reasonable estimate of the prevalence of HIV among study participants' current and potential partners. Certainly, participants who were seronegative at baseline likely differ in their future risk from those who were positive, so screening prevalence may overestimate the risk of exposure during the trial. Alternatively, screening seroprevalence may underestimate the proportion at risk since many of those exposed to HIV never become infected. Ultimately, we choose to set the prior expectation of the proportion at risk equal to the seroprevalence observed at screening but specify a broad distribution that allows for considerable variation from this mean.

HIV prevalence among those screened at each site is given in Table 2.4. In the unstratified model, the prior expectation for the proportion at risk, $P(Risk) = 1 - \exp(-\rho)$, was set equal to HIV prevalence at screening across all sites (an average of the prevalence from each site, weighted by sample size). In the stratified model, the prevalence from each site was used to set the prior for site-specific estimates of ρ. The beta distribution on the proportion at risk was right-skewed, however, to reflect our assumption that the participants enrolled in the trial likely have a lower risk of exposure during the study period.

We chose a constant baseline hazard- $\lambda(t) = \lambda$ in the non-stratified model and $\lambda_j(t) = \lambda_j$

Table 2.4: HIV prevalence at screening and observed seroincidence per 100 person-years (PY).

Site	Prevalence (%)	Seroincidence (100PY)
All Sites	22%	3.4
Hlabisa, South Africa	28%	7.0
Durban, South Africa	22%	5.2
Kamwala, Zambia*	22%	4.1
Blantyre, Malawi	22%	3.1
Harare, Zimbabwe	21%	1.2
Chitungwiza, Zimbabwe	19%	1.6
Lilongwe, Malawi	18%	0.89

Note: (*) Screening data not available; prevalence estimated based on seroincidence.

for all sites j in the stratified model- because we expected that risk of seroconversion was constant over time. Examination of the data confirmed that this assumption was reasonable as no trend in seroincidence over time was apparent. For the non-stratified analysis, we set a conjugate gamma prior on λ. For the site-stratified analysis, we developed a prior for each site's average baseline hazard that was correlated with the proportion at risk in that site. As defined in Section 2.2, the baseline hazard among at-risk participants in site j is a function of the baseline hazard for all participants in site j, λ_j, and the proportion of participants at risk in site j, $1-\exp(-\rho_j)$, and is equal to $\lambda_j/\big(1-\exp(\rho_j)\big)$. In our stratified analysis, we assumed that the probability of exposure to HIV varied across sites due to varying prevalences but that the risk of seroconversion given exposure was roughly similar. To reflect this assumption, we set the expectation of the baseline hazard among at-risk individuals to a single, constant hazard, θ, shared across sites. Hence, the prior assumed $\lambda_j/\big(1-\exp(\rho_j)\big) = \theta$ for all sites j. It follows that the prior distribution for the average baseline hazard for all participants in site j, λ_j, had expectation $\theta\big(1-\exp(-\rho_j)\big)$ for all sites j. Conjugate gamma priors were specified for λ_j accordingly: $\pi(\lambda_j|\rho_j,\theta,c) = Gamma(\,\theta\big(1-\exp(-\rho_j)\big)\times\tau,\,\tau\,)$,

where τ is a rate parameter. A gamma prior was also set on θ, and an inverse gamma prior was set on τ.

The relationship between the proportion at risk and baseline hazard imposed by this prior is illustrated in Figure 2.3, which shows the prior distribution of average baseline incidence, $\lambda_j \times 100$ person-years, for four hypothetical sites. Each site has a different proportion of its population at risk, but the average incidence among those at risk is the same. In this example, $\theta = 40$ per 100 person-years is used as the average seroincidence among those at risk, and the rate parameter, $\tau = 50$, is also equal across sites. We see that the prior distribution for the average baseline hazard among all participants in a site is higher when the proportion of the site's population at risk is larger. The observed relationship makes sense intuitively; when more of the population is exposed but the probability of seroconversion given exposure remains the same, we expect to observe a greater seroincidence.

Finally, the prior for our parameter of primary interest, β, was non-informative. Prior distributions for other parameters were specified in such a way that, under the prior, the expected distribution of incidence in the placebo arm would be consistent with seroincidence previously observed in similar trials.

We set the following priors on model parameters. For the non-stratified model,

$$\text{Prior:} \begin{cases} \beta \sim N(0, 100) \\ \lambda \sim Gamma(0.0625 \times 50, 50) \\ 1 - \exp(-\rho) \sim Beta(2 \times \frac{Prev}{1-Prev},\ 2) \\ \eta \sim Gamma(4, 1) \end{cases}$$

where $Prev$ is the overall seroprevalence from Table 2.4. For the stratified model,

$$\text{Prior:} \begin{cases} \beta \sim N(0, 100) \\ \lambda_j | \rho_j, \theta, c \sim Gamma(\theta\big(1 - \exp(-\rho_j)\big) \times c,\ c) & \text{for each site } j \\ 1 - \exp(-\rho_j) \sim Beta(2 \times \frac{Prev_j}{1-Prev_j},\ 2) & \text{for each site } j \\ \eta_j \sim Gamma(0.5, 2) & \text{for each site } j \\ \theta \sim Gamma(20, 25) \\ \tau \sim InverseGamma(2, 20) \end{cases}$$

where $Prev_j$ is the prevalence for site j, as given in Table 2.4.

Figure 2.4 shows the prior distribution of seroincidence in the placebo arm for the non-stratified and stratified models, produced by 2,500 replications of a simulation that generated event times under both models for 1,000 participants in each of seven study sites. The distribution of expected seroconversion for the stratified model is predictably broader than for the non-stratified model as the former is a mixture of prior distributions from each site. Rates of seroconversion observed in the placebo arms of similar previously published trials are indicated by plotting symbols. We note that the incidence observed in the CAPRISA 004 trial [4] is nearly twice as high as that observed in the other studies. This outlying observation can partly be explained by differences in enrollment criteria. Overall, we see that the distribution of expected seroincidence in the placebo arm under the prior specified is consistent with those previously observed. Also, while the distribution of seroincidence under the prior appears quite informative, we note that prior distributions for individual parameters were specified to allow for considerable variability. As part of model assessment, we compared prior and posterior distributions for all model parameters.

To obtain posterior samples from the non-stratified and stratified compound Poisson frailty models, we repeated the Gibbs sampling algorithm described in Section 2.2.2 for 100,000 iterations. The first 20,000 samples were discarded for burn-in and the remaining samples from parameters' posterior distributions were thinned, keeping every tenth sample. Parallel chains were run from different starting values to confirm that models converged to the same estimates. Sampling was performed in `R` version 3.0.3 with independently written code. Other model diagnostics performed included examining cumulative distribution and trace plots. Posterior distributions for model parameters were visually compared to prior distributions using the `density` function.

We also computed Bayesian p-values for survival at six month intervals using the data generated from the joint posterior predictive distribution. While Bayesian p-values were easily calculated and interpretable measures of model fit for simulations (Section 2.3), the appropriateness and interpretability of Bayesian p-values is questionable in this application because of interval censoring. Specifically, there is not a known event time against which to compare predicted event times. Instead, event times are sampled conditional on the interval in which they occurred and current parameter samples at each iteration of the

	Cox Model		CP Frailty Model	
	Est.	(Interval)	Est.	(Interval)
Non-stratified	0.28	(-0.11, 0.53)	0.31	(-0.13, 0.58)
Stratified	0.27	(-0.12, 0.52)	0.36	(-0.037, 0.62)

Table 2.5: Estimated effectiveness and 95% interval of 0.5% PRO 2000 gel relative to placebo gel in reducing risk of seroconversion for the Cox and compound Poisson (CP) frailty model.

Gibbs sampler. Thus, we are comparing model-based samples of the event times to predicted survival. The Bayesian p-values presented in this application as well as the next chapter's (Section 3.4) should be considered with this limitation in mind.

2.4.2 Results

Incidence of seroconversion for 0.5% PRO 2000 and placebo gel groups are displayed in Figure 2.5. The rate of seroconversion was lower in the treatment group than in the placebo group for all but one site (Durban, South Africa). We also see that the level of seroincidence varies considerably by site, suggesting that allowing the baseline hazard and frailty distribution parameters to vary by site should improve model estimation.

Table 2.5 reports estimates of the population- and individual-level effectiveness of 0.5% PRO 2000 gel in reducing risk of seroconversion in comparison to the placebo gel from the Cox model and the proposed compound Poisson (CP) frailty model, respectively. Without stratification, the individual-level effectiveness estimate from the compound Poisson frailty model was 0.31 (95% HPDI: -0.13, 0.58), compared to the effectiveness estimate from the Cox model, 0.28 (95% CI: -0.11, 0.53). Under the stratified compound Poisson frailty model, 0.5% PRO 2000 gel was estimated to be 36% effective in reducing the risk of seroconversion for an exposed individual (95% HPDI: -0.037, 0.62). In comparison, the stratified Cox model estimate of the effect of the intervention in a population was 27% (95% CI: -0.12, 0.53).

Site-stratification had minimal impact on Cox model estimates; it is likely that the non-stratified Cox model is relatively robust to site differences because of the flexibility

of the baseline hazard. Site-stratification did affect results for the proposed model: the estimated effectiveness from the stratified compound Poisson frailty model was higher than the estimated effectiveness from the non-stratified model. Since both the stratified Cox and compound Poisson frailty models allowed the baseline hazard to vary by site, it is likely that allowing for the proportion at risk to vary by site accounts for this difference, which would be consistent with theory (Section 1.1) and simulation results (Section 2.3). As seen in Figure 2.6, there was considerable variation in the estimated proportion at risk across sites.

Figure 2.6 shows site-specific estimates from the stratified compound Poisson frailty model of the baseline seroincidence, $\lambda_j \times 100$ person-years, and proportion at risk, $1 - \exp(-\rho_j)$, plotted against the seroincidence observed in the placebo arm. Unsurprisingly, estimates of the proportion at risk and the baseline incidence for each site were strongly correlated with the observed seroincidence. Moreover, the HPD intervals around the non-stratified estimates of both parameters actually exclude some of the site-specific estimates or even, as is the case with estimated baseline seroincidence in Lilongwe and Hlabisa, the entire site-specific interval.

Figure 2.7 shows Kaplan-Meier survival for the observed data (bold line) as well as survival curves from the posterior predictive distribution (gray lines). Bayesian p-values at each six month interval are also given. Plots on the left side are from the unstratified model while those from the stratified model are shown on the right. Separate plots are given for all study arms combined, the placebo gel arm, and the 0.5% PRO 2000 gel arm (from top to bottom). The midpoint between last negative and first positive test was used as the event time for the observed data. 95% confidence intervals are also given for the observed survival (dashed lines). Bayesian p-values were calculated using every tenth iteration of the Gibb's sampler (after burn-in), but survival curves are only shown for every 250th iteration.

For both plots, we see that survival is underestimated in the first 6 months and the Bayesian p-values for that interval are close to one. This discrepancy between the observed and predicted survival is not surprising as there is a lag of several weeks between HIV infection and seroconversion, and all participants who were HIV-infected at baseline were removed from the analysis. The constant baseline hazard assumed by our model would not

have been able to accommodate this time-dependence in the rate of events. This initial overestimation of risk is generally balanced out over the remainder of follow-up, as the majority of those p-values are below 0.5. Overall, both the unstratified and stratified models appear to provide a good fit to the data.

Plots comparing prior and posterior distributions for all model parameters in both the non-stratified and stratified analyses are given in Figures 2.8- 2.13. We observe that parameters' posteriors varied considerably from their priors, with the exception of η. In the unstratified analysis, the posterior for η (lower-right panel of Figure 2.8) appears nearly identical to its prior distribution. In the stratified analysis, site-specific posterior distributions of η (Figure 2.13) are also similar to the prior distributions. The posterior distributions of η for Durban and Hlabisa (second row), the two sites with the highest seroincidence, only show slight variation from the prior distribution. We see in Figures 2.11 and 2.12 that the posteriors for site-specific baseline hazard and proportion at risk vary in both location and spread from one another.

Finally, we see that the posterior distribution for the proportion at risk in the unstratified model (Figure 2.8) suggests poor mixing. Examination of trace plots for this parameter indicates that sampling was concentrated in a small range (for the starting seed shown, 5-15% at risk) before jumping to another area (45-75%). This behavior was observed at multiple starting seeds for the Gibbs sampler and persisted even in longer chains. By comparison, site-specific distributions of the proportion at risk in the stratified model do not show this problem. These posteriors generally show more variability than their respective prior distributions, and trace plots indicate convergence.

2.5 Discussion

We have presented a Bayesian hierarchical model that adjusts for heterogeneity in risk when estimating the individual-level effect of a covariate on survival. Unlike the gamma distribution, which is more commonly chosen for frailty modeling, the compound Poisson distribution reflects the mechanisms of HIV risk and allows some participants to have no exposure to HIV and, therefore, no risk of seroconversion. The hierarchical definition of the compound Poisson distribution presented here is intuitive, and its parameters are in-

terpretable. As such, the proposed Bayesian model is a useful alternative to frequentist analogs. Frequentist formulations of the compound Poisson frailty model have relied on the Laplace transformation of the compound Poisson density, the parameters of which have no meaningful interpretation. Furthermore, a computational strength of our model is that it exploits hierarchical modeling and Bayesian estimation methods to avoid approximating an infinite series. The density of the continuous part of a compound Poisson random variable is not available in closed form, which burdens frequentist model estimation. In our sampling procedure, we sample the frailty term conditional on the underlying Poisson random variable, and this distribution is available in closed form.

A disadvantage of the proposed model, as well as all current frailty models for non-clustered univariate survival data, is that heterogeneity is modeled with a single frailty distribution and only information available in event time and study arm are used to sample individual's frailties. While additional data collected on participant's risk factors could improve heterogeneity modeling, existing methods provide no mechanism for using risk-related covariates to inform individuals' frailties. We address this problem in the next chapter.

The approach proposed also assumes that a participant's frailty is constant over time. In the case of HPTN 035, this assumption may be reasonable, since the majority of participants remained in the same, sole partnership throughout the study. In general, allowing participants' frailties to change over time in order to reflect a change in partnerships or behavior would likely improve heterogeneity modeling. To do so would present challenges, however, as current time-varying frailty methods are quite limited for univariate survival data.

Despite these limitations, we have demonstrated that the inclusion of a compound Poisson frailty term into a survival model is able to capture some of the heterogeneity in HIV risk. When applied to data from HPTN 035, the site-stratified compound Poisson frailty model estimated that 0.5% PRO 2000 gel is 36% effective in reducing risk of seroconversion for exposed individuals in comparison to placebo gel. This estimate was larger than the Cox model estimate of effectiveness.

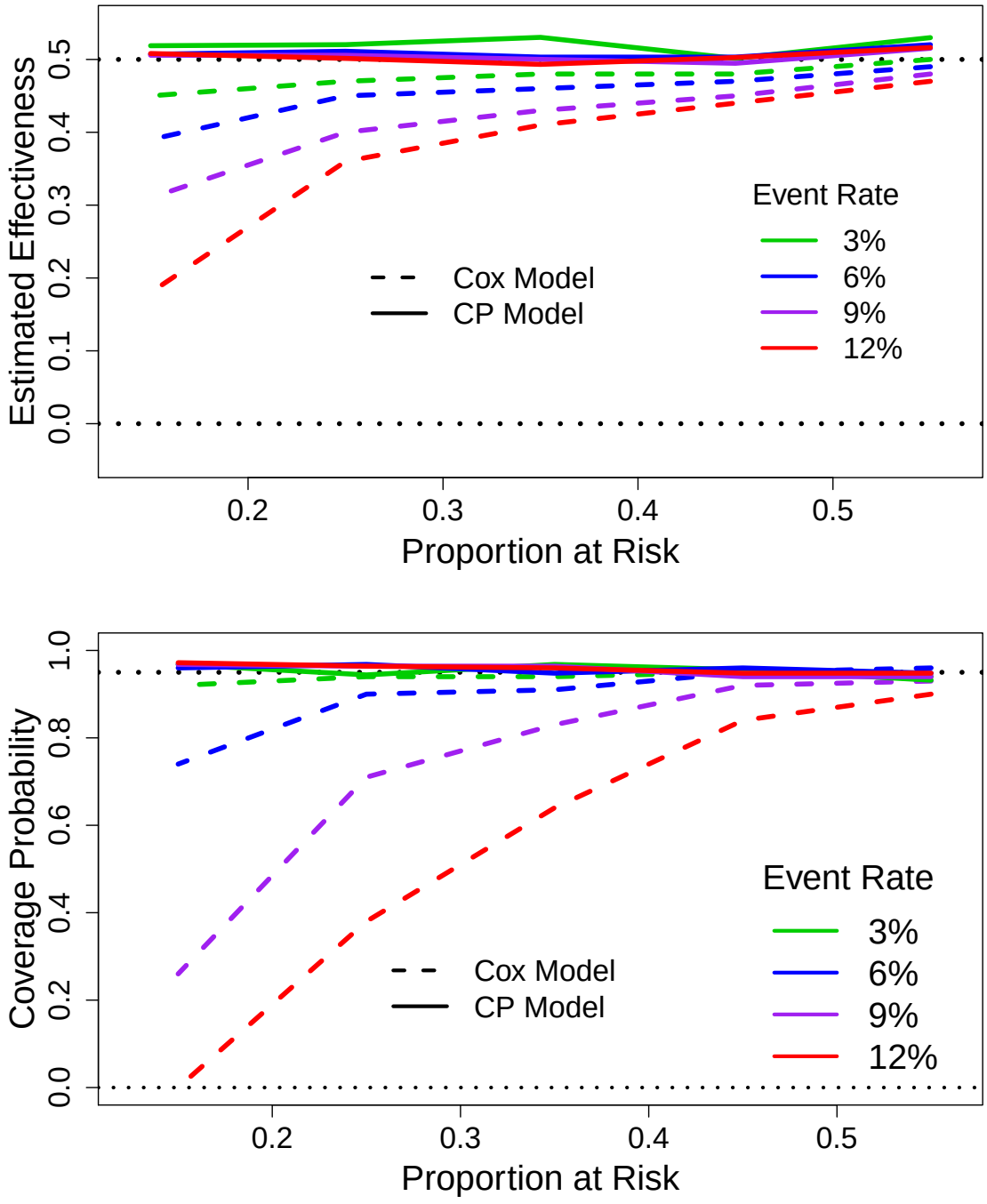

Figure 2.1: Estimated effectiveness (upper panel) and coverage of 95% intervals (lower panel) from simulation studies.

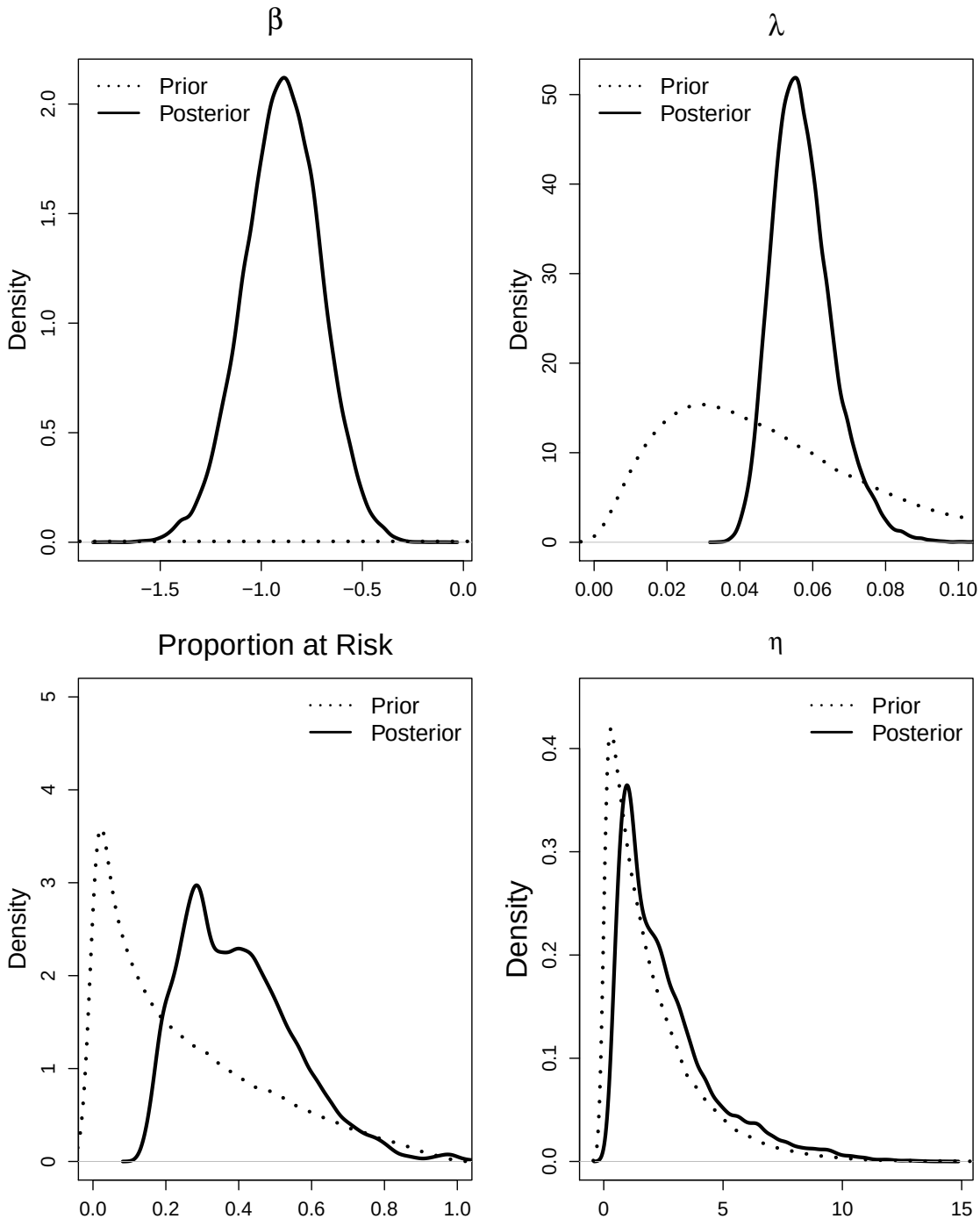

Figure 2.2: Prior and posterior distributions for model parameters from a single replication of simulation using priors in Equation (2.7). Data generated with 25% proportion at risk and a 6% event rate.

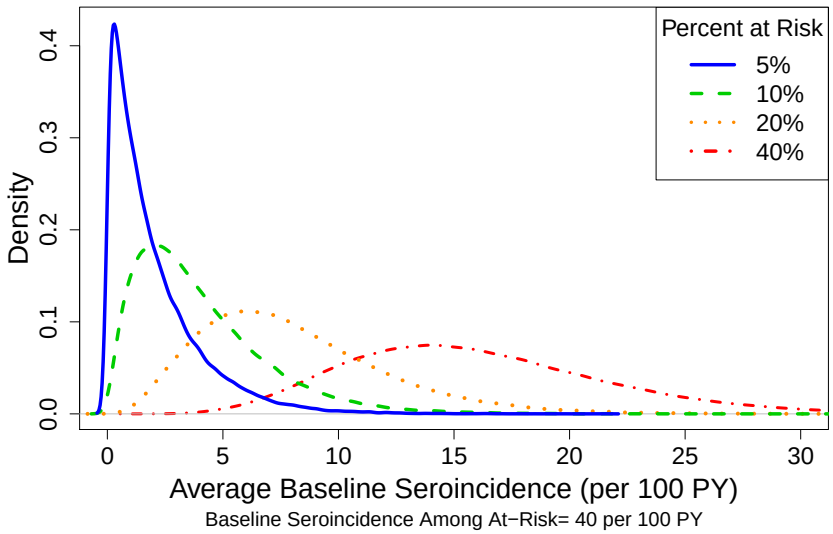

Figure 2.3: Prior distribution of average baseline incidence when proportion at risk varies and average seroincidence among those at risk is equal across all sites.

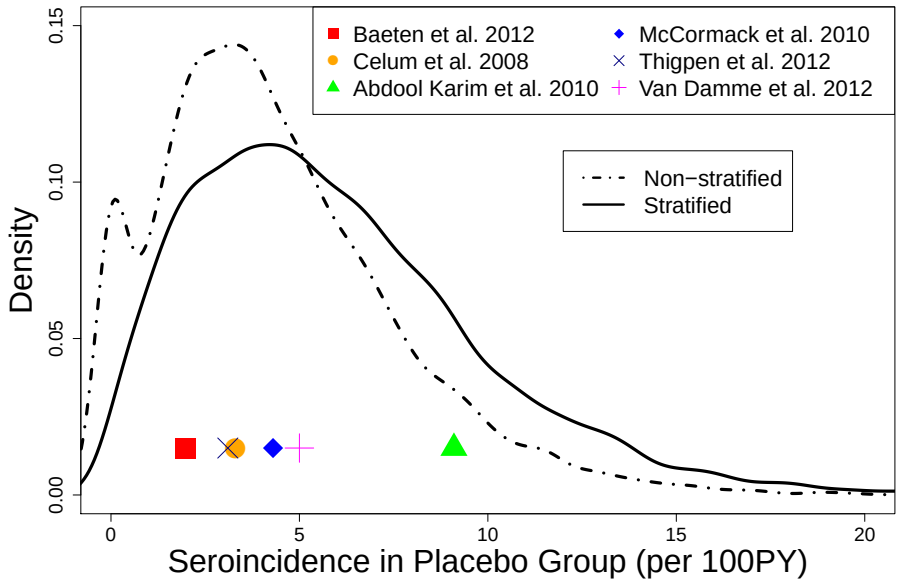

Figure 2.4: Density of seroincidence in placebo arm under specified priors.

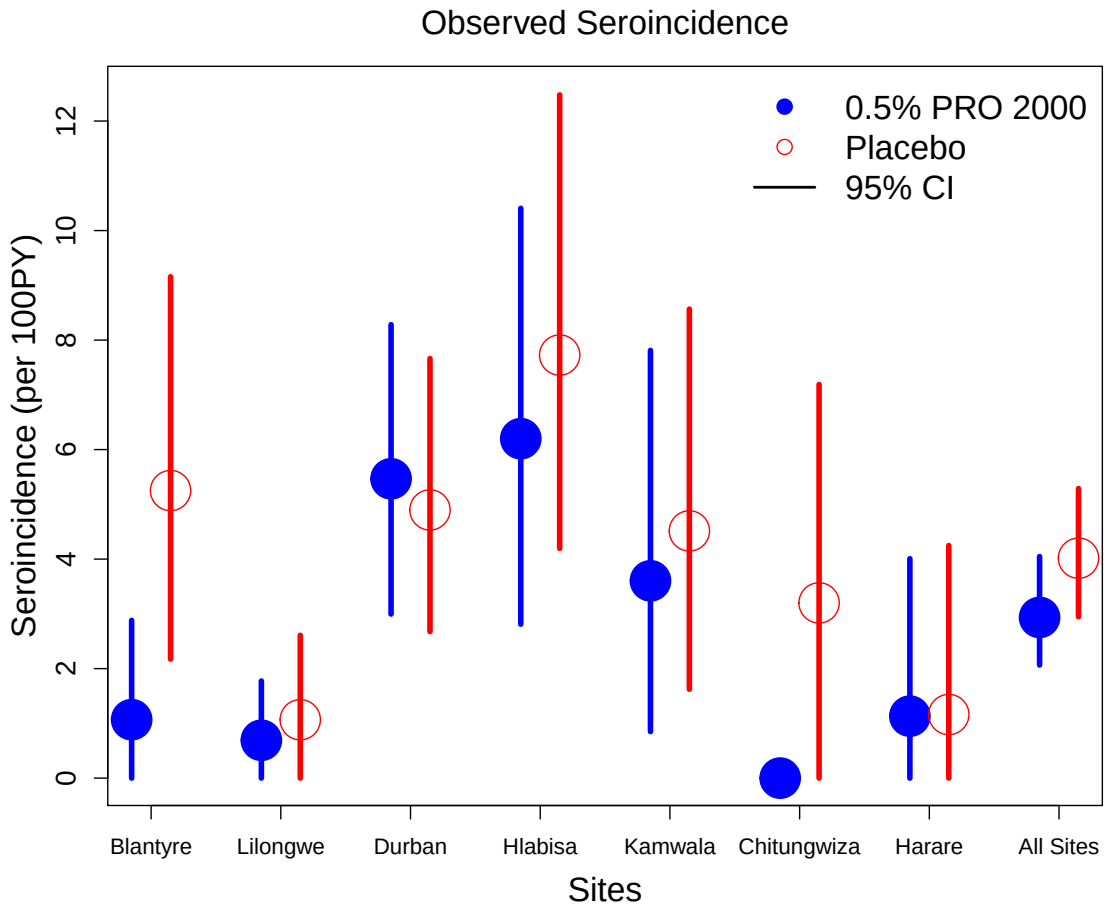

Figure 2.5: Observed seroincidence for 0.5% PRO 2000 and placebo gel groups for each site in HPTN 035.

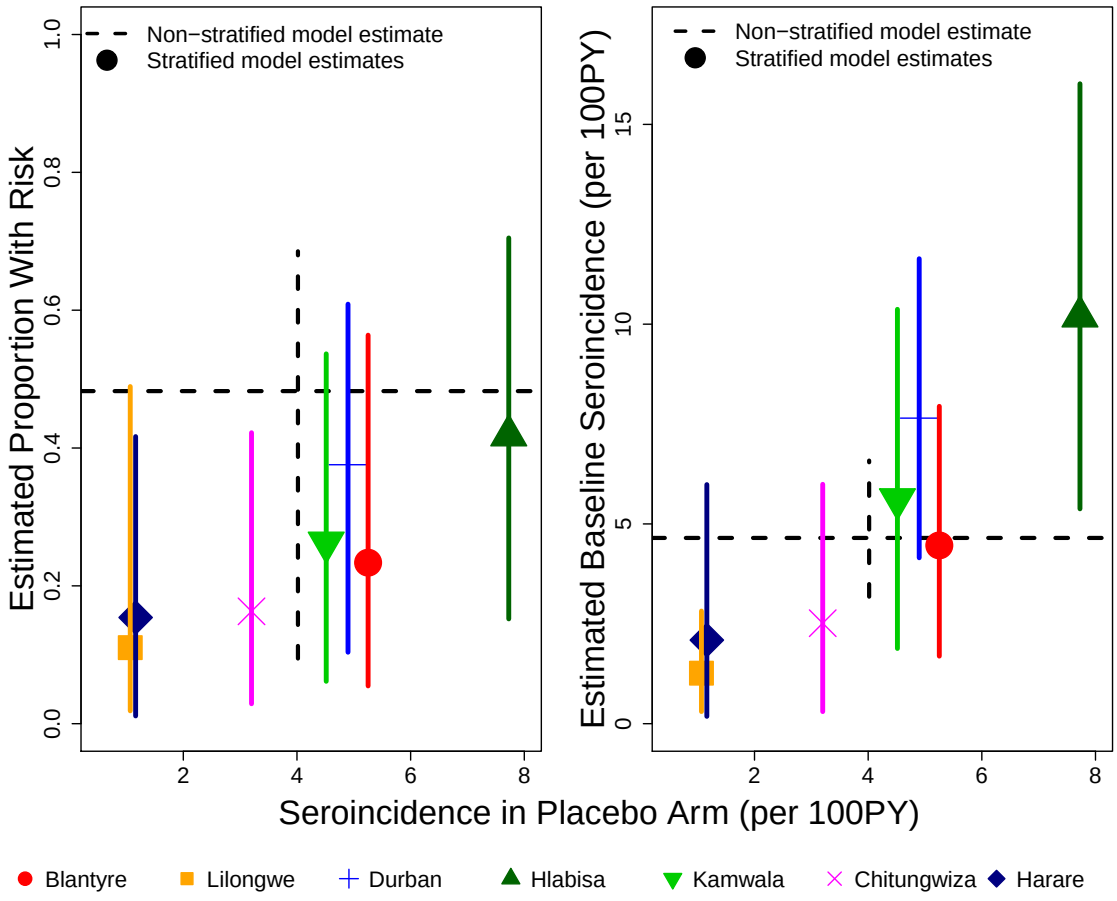

Figure 2.6: Site-specific estimates of proportion of participants at risk of seroconversion, $1-\exp(-\rho_j)$, (left panel), and baseline seroincidence, $\lambda_j \times 100$, (right panel) versus observed seroincidence in placebo arm, represented by plotting symbols. Solid lines depict 95% HPD intervals around site-specific estimates. Estimates and intervals from non-stratified model are represented by horizontal and vertical, respectively, dashed lines.

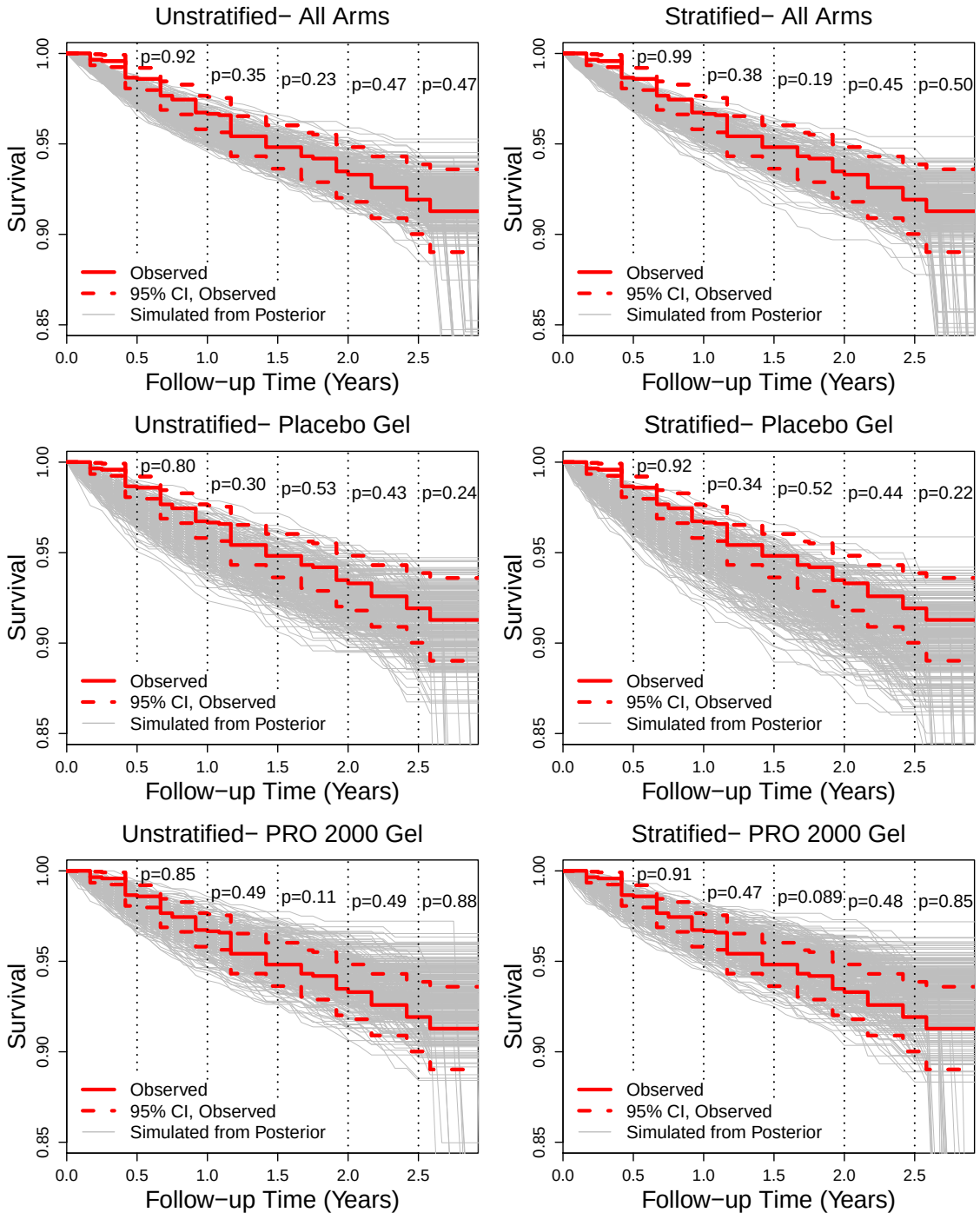

Figure 2.7: Kaplan Meier curves for observed data (bold, solid lines) and surrounding 95% confidence intervals (dashed lines) with survival curves generated the posterior predictive distribution (gray lines). Bayesian p-values comparing observed to predicted survival are given at six-month intervals.

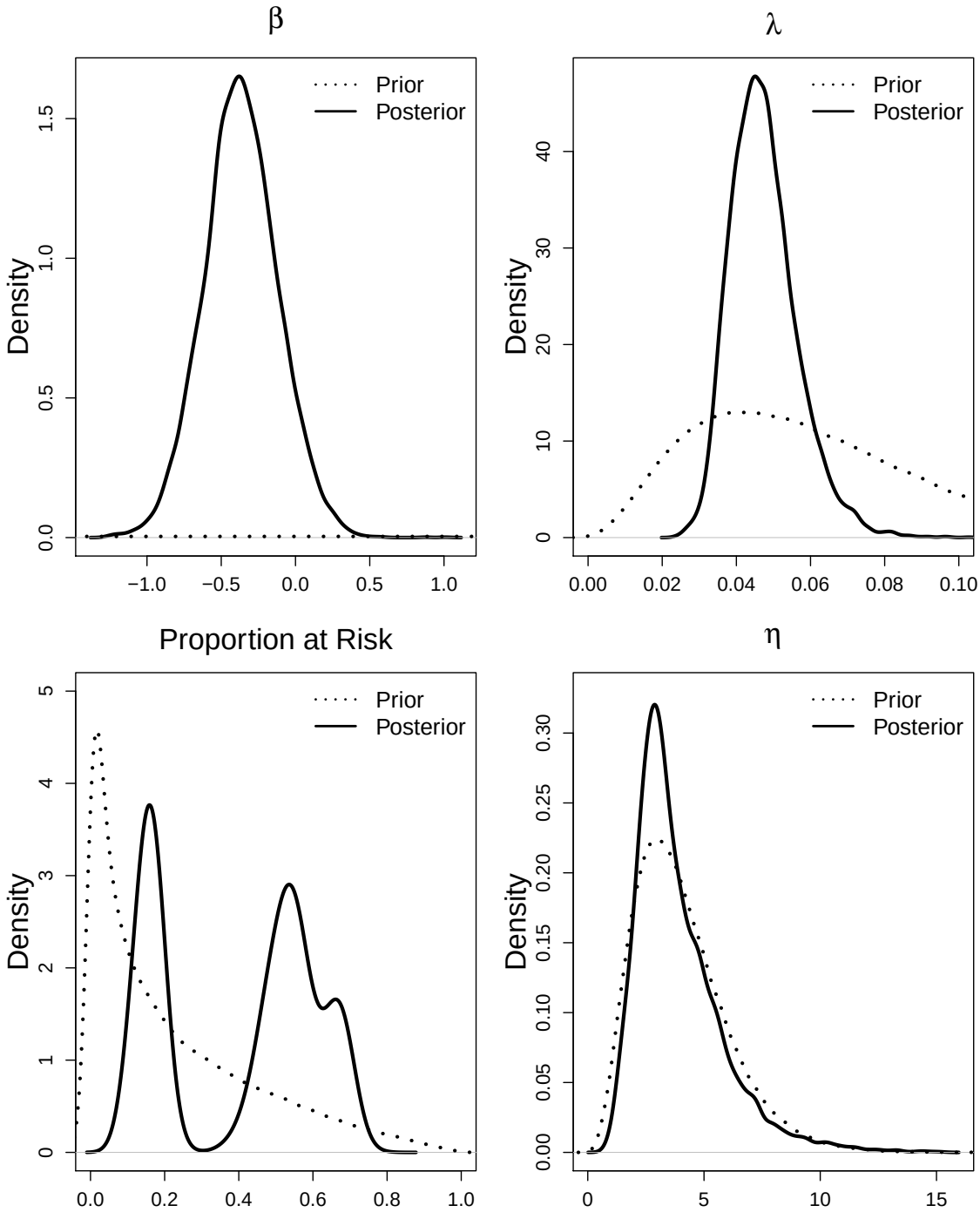

Figure 2.8: Prior and posterior distributions of model parameters from non-stratified analysis of HPTN 035 data.

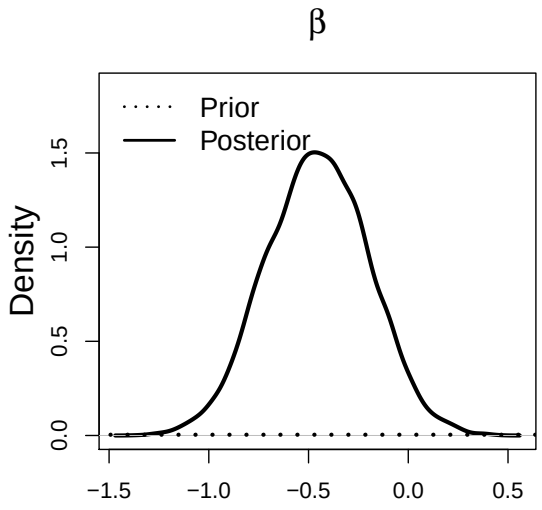

Figure 2.9: Prior and posterior distribution of log-HR, β, from stratified analysis of HPTN 035 data.

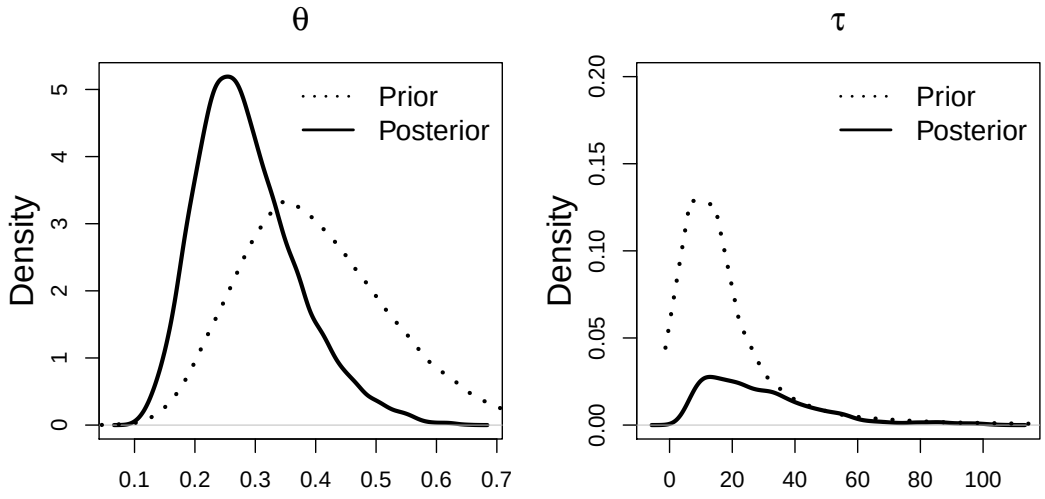

Figure 2.10: Prior and posterior distributions of hyperparameters θ and τ from stratified analysis of HPTN 035 data.

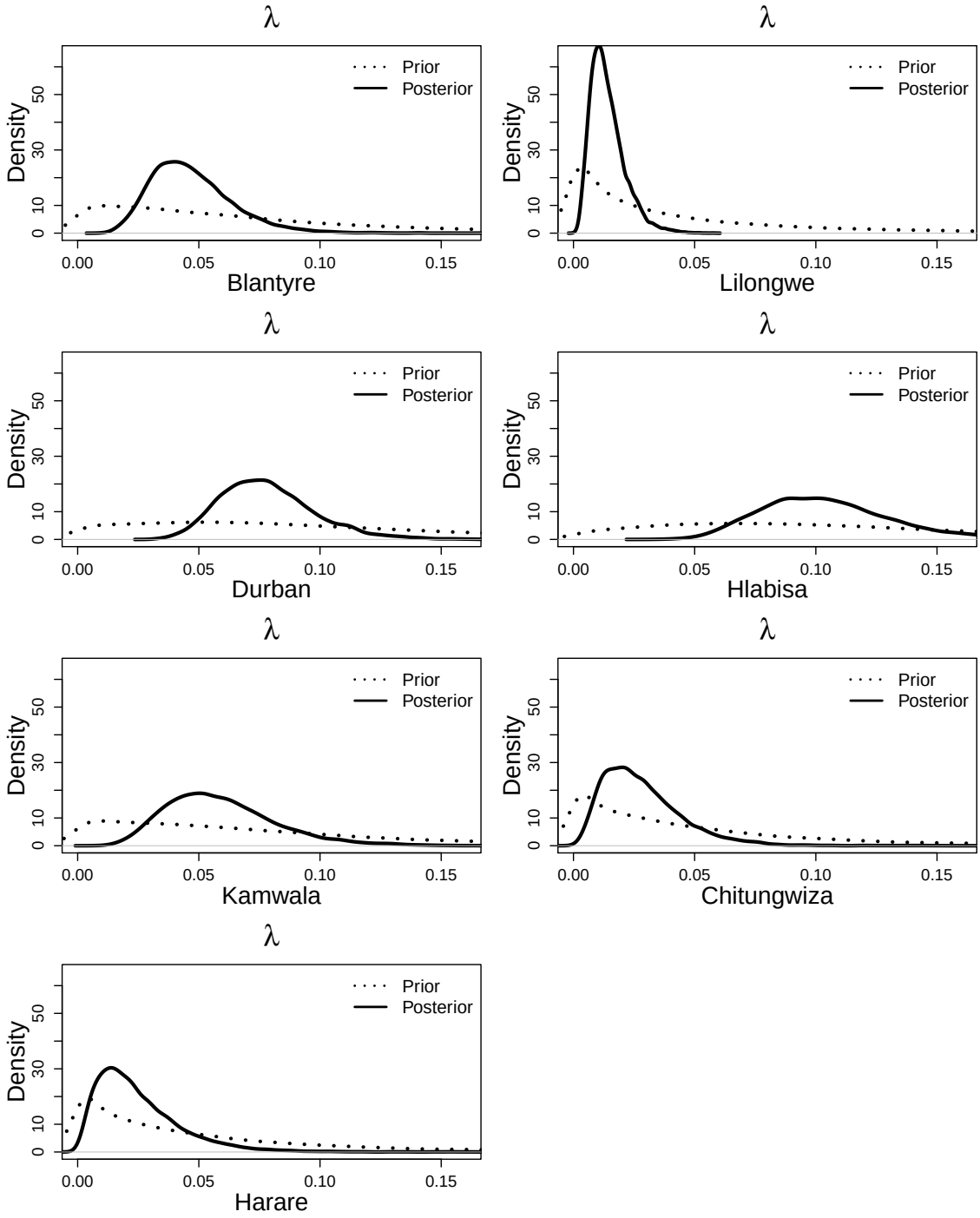

Figure 2.11: Prior and posterior distributions of site-specific baseline hazards, λ_j, from stratified analysis of HPTN 035 data.

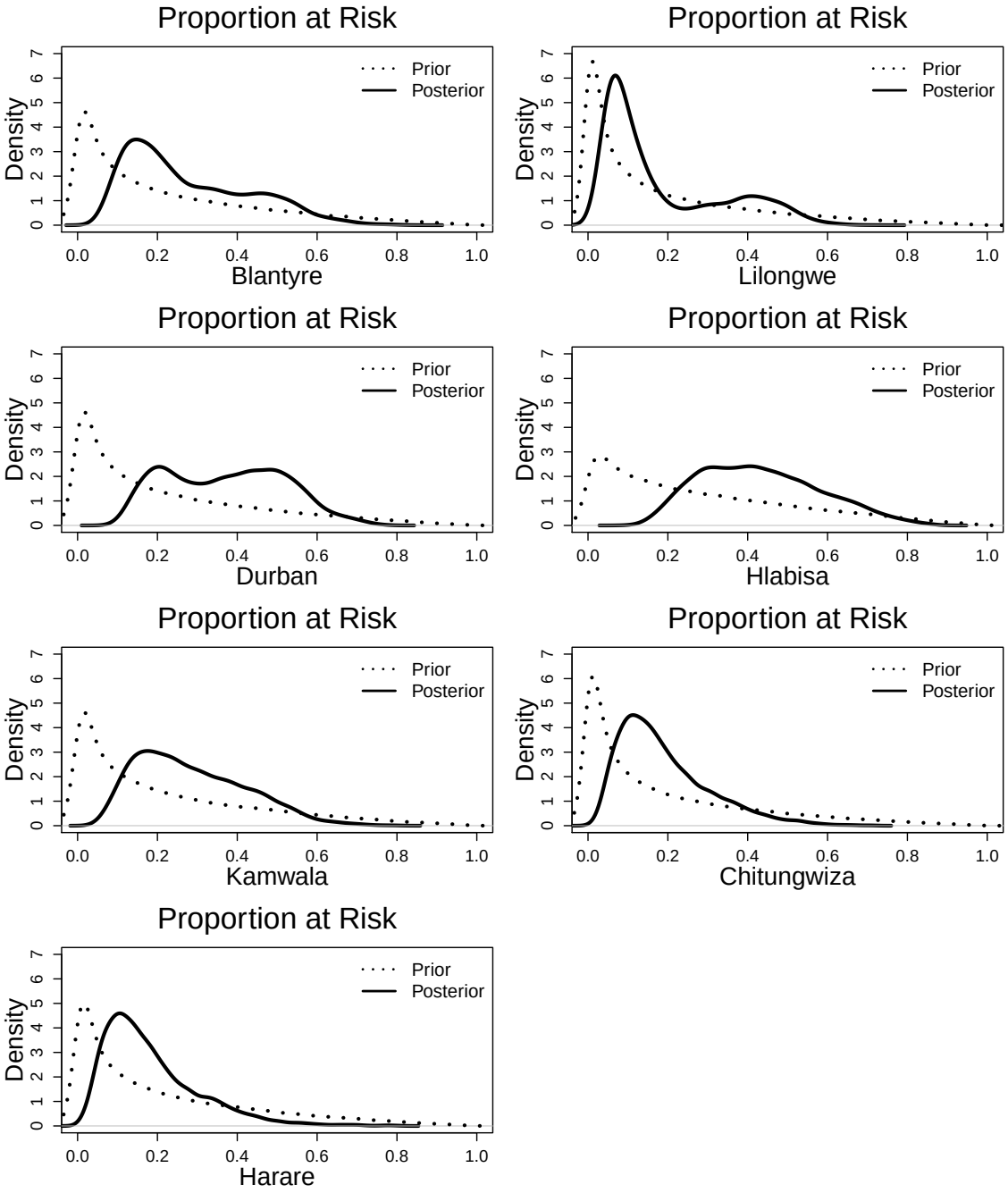

Figure 2.12: Prior and posterior distributions of the site-specific proportion at risk, $1 - \exp(\rho)$, from stratified analysis of HPTN 035 data.

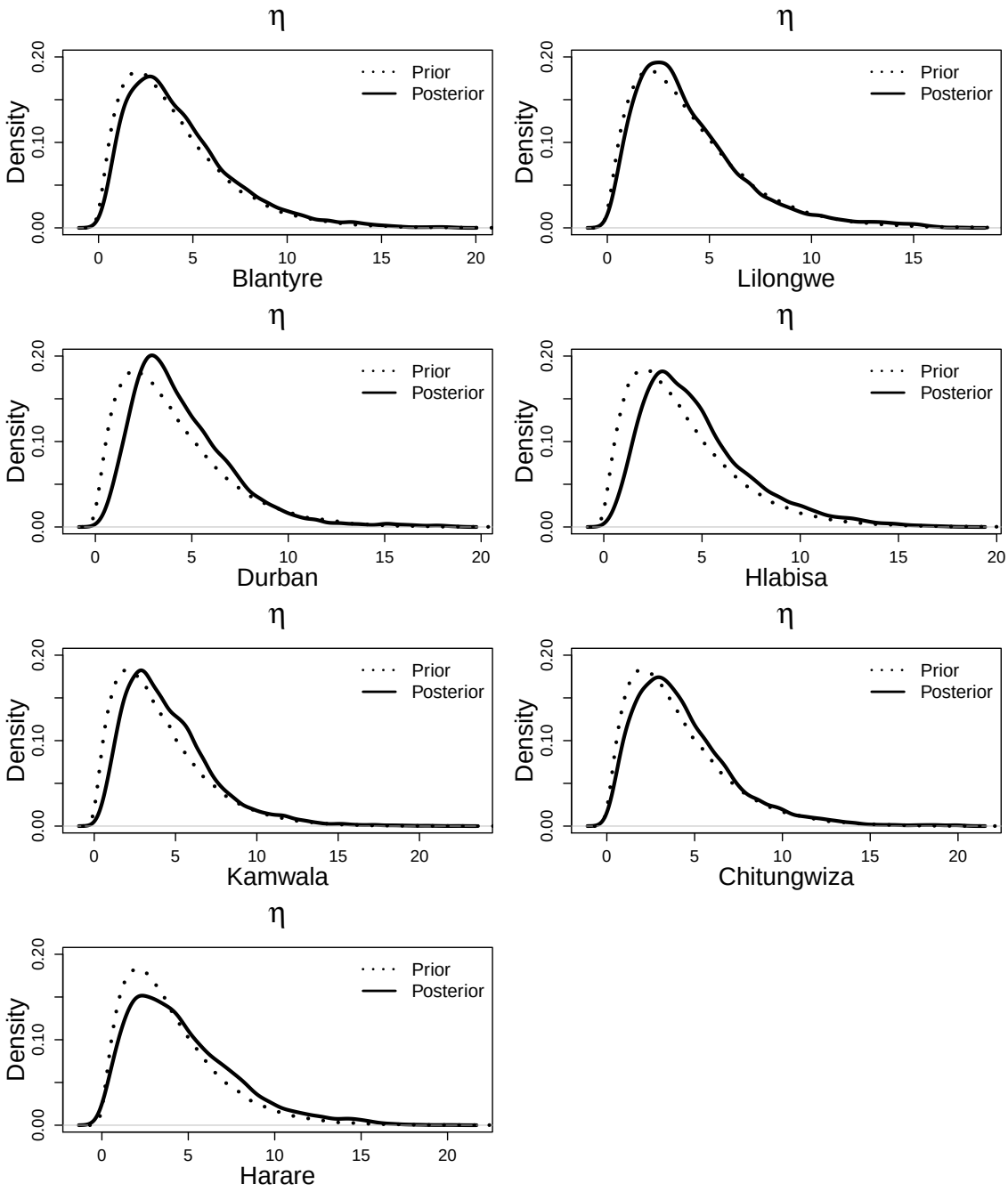

Figure 2.13: Prior and posterior distributions of site-specific η from stratified analysis of HPTN 035 data.

Chapter 3

LATENT CLASS APPROACH TO COMPOUND POISSON MIXTURE FRAILTY MODEL

3.1 Introduction

In studies where the outcome is time-to-event, the Cox proportional-hazards model is often used to estimate a population-averaged hazard ratio associated with an intervention or exposure of interest [18]. As discussed earlier, when heterogeneity in risk is present, the assumption of proportionality does not hold, and the population-level estimate is a biased estimate of the change in risk for an individual. In order to account for heterogeneity, risk-related covariates can be included in the Cox regression model, but this approach requires that all risk factors are known and measured without error and that they have a proportional impact on the population level. An alternative method is frailty modeling, which uses a random effect added to a survival model to adjust for unobserved heterogeneity [50]. Frailty model estimation requires specification of a mixing, or frailty, distribution. Individuals' frailties are assumed to vary according to this distribution.

A limitation of current frailty methods for univariate survival data is that they provide no mechanism for using risk-related covariates to inform heterogeneity modeling. While covariates can be included in the Cox regression model, they are subject to the limitations discussed previously. When covariates affecting risk are unobserved or the assumption of proportionality is inappropriate, we may want to use available covariate data to model frailty.

In this chapter, we propose a novel model that adjusts for heterogeneity in risk by varying frailty distributions across latent risk classes modeled by participant covariates, resulting in a mixture of frailty distributions. Via an ordinal latent class regression model, groups of participants with similar risk characteristics are identified. Individuals within a class share a frailty distribution, and their frailties vary accordingly. In this way, risk-

related covariates do not determine an individual's frailty but rather specify its distribution. Meanwhile, frailty distribution parameters vary systematically across classes to reflect their order. That is, frailty distributions for classes in which participants have more risk factors are specified such that individuals within that class are more likely to have higher risk i.e. higher frailty. While other classification methods could be used, ordinal latent class regression is particularly appealing in contexts where we are interested in the resulting risk classifications, in addition to accounting for heterogeneity.

The proposed model is similar to some approaches for clustered survival data that allow frailty distributions to vary across clusters. For example, Moger et al. (2004) and Moger and Aalen (2005) have developed a frailty model where frailty distribution parameters vary across clusters defined by family; frailty for individuals within a family vary according to that shared distribution. Frailty methods for clustered data are also often used in the analysis of time-to-event outcomes from multi-center clinical trials in order to model heterogeneity across sites. These methods are limited in their applicability because, for many studies, observations are not clustered on a characteristic, like family, that explains the majority of unobserved heterogeneity. Furthermore, in some multi-center trials, site may be only one of several factors influencing an individual's risk, and solely specifying separate frailty distributions for each site may not sufficiently account for unobserved heterogeneity. With this motivation in mind, we have developed a mixture frailty model that, instead of relying on observable clusters, models latent classes of participants with similar probabilities of exposure and levels of risk using covariate data.

For the proposed mixture frailty model, we select a frailty distribution that reflects the sources of heterogeneity in risk. The compound Poisson distribution is used to model frailty for all latent classes as this distribution models a risk mechanism common in medical contexts where an individual's risk of an event is the result of independent, competing exposure processes. Examples include patterns of risk behavior resulting in HIV infection, atherosclerotic lesions causing a cardiovascular event, and tumors leading to cancer recurrence. In Chapter 2, we presented a hierarchical definition of the compound Poisson distribution that reflects this risk mechanism. The number of exposure processes for each individual, frequently either measured with error or unknown, is modeled as a Poisson random variable,

which allows for a participant to have no risk (i.e. no exposure processes). Each exposure process is assumed to present a constant level of risk over time, which is modeled as a Gamma random variable. Separate exposure processes pose independent, competing risks; given exposure, one of the processes eventually leads to an event. An individual's risk of an event during a specified time period is the sum of these competing risks. Following this hierarchical definition of the compound Poisson distribution, frailty distributions for each latent risk class will be specified by two parameters: each class will have a different mean number of exposure processes and different shape parameter for the gamma random variable modeling risk associated with each process. Frailty distributions also reflect class ordering: individuals in a class with more risk factors will have a larger number of expected exposure processes as well as larger expected risk for each process. This model allows for some participants, even those in higher risk classes, to have no risk of an event.

While similar to a cure rate model, the compound Poisson mixture frailty model has multiple advantages over such a model. Cure rate models are used to account for heterogeneity when a fraction of the population is not at risk of an event, or cured, and the probability of cure can be modeled with risk-related covariates. Non-frailty cure rate models do not account for heterogeneity among those who are at risk, and frailty cure-rate models are subject to the limitations of frailty models for univariate survival data discussed above. In the proposed model, participant risk factors influence both the probability and the level of exposure via classification in risk groups that share frailty distribution parameters. In particular, the parameter dictating the probability of risk for an individual within each class, the mean parameter for the Poisson distribution, also influences the distribution of risk among those who are at risk. This relationship is intuitively appealing: we frequently expect that many factors that increase an individual's risk of exposure also increase its magnitude. The model we present here may better reflect the risk-generating mechanism in many medical contexts.

3.2 Methods

We have developed a mixture frailty model that uses risk-related covariate information to model heterogeneity in risk. In this section, we first present a latent class approach for

identifying groups of individuals with similar risk profiles. Next, we provide a hierarchical definition of the compound Poisson distribution. Frailties for individuals within a latent class follow a shared compound Poisson distribution, such that frailties for all individuals, across all classes, follows a compound Poisson mixture distribution. Then, the compound Poisson mixture-distributed frailty is incorporated into a survival model. Finally, we discuss Bayesian estimation procedures for this model, including consideration of prior elicitation.

3.2.1 Latent Class Model

We define an ordinal latent class regression model to identify groups of participants with similar characteristics influencing risk of an event [36]. Classification for an individual follows a multinomial distribution and depends on risk-related covariates through a cumulative link model [7]. In this section, we present a general cumulative link model for latent classification, which models classification probabilities separately for pairs of adjacent classes.

Let R_i denote the latent class response for participant i in one of K ordered categories. Latent classes are ordered such that those participants in class $k = 1$ have characteristics associated with the lowest levels of risk and those in class K have the highest risk. Let the probability of class membership for participant i in class k be indicated as $\phi_k(\mathbf{w}_i) = P(R_i = k|\mathbf{w}_i)$, $k = 1, \ldots, K$, where $\mathbf{w}_i$ is a vector of covariates for participant i and $\sum_{k=1}^{K} \phi_k(\mathbf{w}_i) = 1$ for all $i = 1, \ldots, n$. Classification for each participant i follows a multinomial distribution, $R_i|\mathbf{w}_i \sim Mult_K\big(\phi_1(\mathbf{w}_i), \phi_2(\mathbf{w}_i), \ldots \phi_K(\mathbf{w}_i)\big)$. In the cumulative link model, the probabilities for this distribution are defined cumulatively, that is, the cumulative probability for membership in class k or below can be expressed as:

$$P(R_i \leq k|\mathbf{w}_i) = \sum_{k^\dagger=1}^{k} \phi_{k^\dagger}(\mathbf{w}_i).$$

It follows that the probability of classification in class k for participant i, $\phi_k(\mathbf{w}_i)$, is the difference $P(R_i \leq k|\mathbf{w}_i) - P(R_i \leq (k-1)|\mathbf{w}_i)$.

Cumulative probabilities for classification are regressed on risk-related covariates via a link function, G^{-1}, the inverse of a continuous cumulative distribution function (cdf) G.

The cumulative link model is

$$G^{-1}\big(P(R_i \leq k|\mathbf{w}_i)\big) = \alpha_k - \mathbf{w}_i\boldsymbol{\gamma}_k \tag{3.1}$$

where $\boldsymbol{\gamma}_k$ is a D- length vector of parameters common to all subjects and α_k is an intercept parameter which satisfies $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{K-1} < \alpha_K = \infty$. It follows that

$$\phi_k(\mathbf{w}_i) = G\big(\alpha_k - \mathbf{w}_i\boldsymbol{\gamma}_k\big) - G\big(\alpha_{k-1} - \mathbf{w}_i\boldsymbol{\gamma}_{k-1}\big),$$

and the parameters $\alpha_{k-1}, \alpha_k, \gamma_{k-1}$, and γ_k must be constrained such that $G\big(\alpha_k - \mathbf{w}_i\boldsymbol{\gamma}_k\big) > G\big(\alpha_{k-1} - \mathbf{w}_i\boldsymbol{\gamma}_{k-1}\big)$ for all $k = 1, \ldots, K-1$. Note that probability of classification in the lowest and highest risk classes, $\phi_1(\mathbf{w}_i)$ and $\phi_K(\mathbf{w}_i)$, respectively, are defined as:

$$\begin{aligned} \phi_1(\mathbf{w}_i) &= G\big(\alpha_1 - \mathbf{w}_i\boldsymbol{\gamma}_1\big) \\ \phi_K(\mathbf{w}_i) &= 1 - G\big(\alpha_{K-1} - \mathbf{w}_i\boldsymbol{\gamma}_{K-1}\big) \end{aligned}$$

since $\alpha_0 = -\infty$ and $\alpha_K = \infty$.

There are several common options for the cdf, G, including the logistic, normal, and extreme value distribution functions. The choice of cdf depends on the application and expected distribution of the linear term on the right-hand side of Equation (3.1). The logistic cdf is the most common choice [7]; using the logit link function, coefficients can be interpreted as log-odds. In a cumulative logit model with predictor w, a unit increase in the value of covariate w is associated with an $\exp(-\gamma_k)$ times increase in the odds of membership in class k or below for all $k = 1, \ldots, K-1$. If γ_k is negative, then $\exp(-\gamma_k)$ is greater than one, indicating that w is associated with increased odds of being in class k or below, i.e. is associated with lower risk. Likewise, if γ_k is positive, then $\exp(-\gamma_k)$ is less than one, indicating that w is associated with higher risk. Use of the logistic cumulative link model is demonstrated in Section 3.4.

3.2.2 Hierarchical Compound Poisson Model

Individuals within a latent class share the same frailty distribution. Because the compound Poisson distribution allows that some participants have no risk of an event and reflects a biological mechanism for risk that arises in many applications, we use it here. In this model,

an individual's overall risk is the result of independent, competing exposure processes that each pose varying levels of risk. We expect the number of exposure processes and risk associated with each to be, on average, lower for participants with fewer risk factors and higher among those with more risk factors. So, we vary these parameters across latent classes, resulting in a mixture of compound Poisson distributions.

The number of exposure processes, N_i, for participant i in class k follows a Poisson distribution with mean ρ_k such that

$$N_i |\, R_i = k \sim Poisson(\rho_k) \qquad k = 1, \ldots, K,\, i = 1, \ldots, n$$

where ρ_k denotes the mean parameter shared by all individuals in class k. This model allows that some individuals have no exposure processes, $N_i = 0$, and thus, no risk of an event and, additionally, that this likelihood of exposure varies by class. For those individuals with $N_i > 0$, each exposure process presents a varying level of risk. The mth exposure process for participant i in class k is associated with a hazard for an event, $\lambda(t)U_{im}$, where $\lambda(t)$ is the average baseline hazard function among all individuals,

$$U_{im} |\, R_i = k \sim Gamma(\eta_k, \nu), \qquad m = 1, \ldots, N_i,\, k = 1, \ldots, K,\, i = 1, \ldots, n,$$

η_k denotes a shape parameter shared by all individuals in class k, and $Gamma(\eta, \nu)$ is a gamma distribution with shape parameter η and rate parameter ν. The scale parameter ν is shared by all individuals. Assuming the risk of an event is additive across exposure processes (i.e. they are independent competing risks), a participant's risk of an event at time t is a product of the baseline hazard, $\lambda(t)$, and the sum of risks associated with each exposure process, which equals $\lambda(t) \sum_{m=1}^{N_i} U_{im}$. We define Z_i, the frailty term for participant i, as follows:

$$Z_i |\, N_i,\, R_i = k \;\; = \;\; \begin{cases} \sum_{m=1}^{N_i} U_{im} \sim \text{Gamma}(\text{shape} = N_i\eta_k, \text{rate} = \nu) & \text{if } N_i > 0 \\ 0 & \text{if } N_i = 0 \end{cases}$$

for all classes $k = 1, \ldots, K$ and participants $i = 1, \ldots, n$.

Since the latent classes are ordered with respect to level of risk, we constrain the parameter values of the compound Poisson distribution for each class to reflect that ordering:

$\rho_1 < \rho_2 < \cdots < \rho_K$ and $\eta_1 < \eta_2 < \cdots < \eta_K$. The resulting compound Poisson distributions are stochastically ordered, i.e. for Z_i, the frailty term representing level of risk for individual i, $P(Z_i > z|R_i = k) > P(Z_i > z|R_i = k-1)$. This framework allows for the possibility that an individual in a lower risk class has a larger risk than some individuals in a higher risk class and vice-versa while still sharing information across classes through the ordering of parameters.

For model identifiability, it is necessary that the expectation of a randomly selected Z_i be equal to one. As the expectation of $Z_i|R_i = k$ is equal to $(\rho_k\eta_k)/\nu$ for participant i, the expected frailty marginalized over all classes, $E(\,(\rho_k\eta_k)/\nu)$, equals $\sum_{k=1}^{K} \frac{\rho_k\eta_k}{\nu}\,\phi_k(\mathbf{w}_i)$. It follows that, for all individuals, the expectation of a randomly selected frailty is set equal to one by defining ν as a function of all ρ_k, η_k, and $\phi_k(\mathbf{w}_i)$:

$$\sum_{k=1}^{K} \rho_k\eta_k \,\frac{1}{n}\sum_{i=1}^{n} \phi_k(\mathbf{w}_i). \tag{3.2}$$

3.2.3 Hazard Regression with Frailty

Having defined a model for frailty conditional on latent classification, we next incorporate frailty into a survival model. Z_i is included in a hazard function for a proportional hazards model as a frailty term which scales the hazard for individual i, as appropriate. We define the ith participant's hazard function as:

$$\lambda_i(t; x_i) = \lambda(t)\, Z_i \exp(\mathbf{x}_i'\boldsymbol{\beta})$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)$ is a P-length vector of the log-hazard ratio associated with a unit increase in covariates $x_1, \ldots, x_P$, respectively. The vector $\mathbf{x}_i = (x_{1i}, \ldots, x_{Pi})$ denotes observed values of these covariates for individual i, which may include an indicator of receiving an intervention or having a condition or some measure of an explanatory variable whose association with the risk of an event is of interest. The baseline hazard function, $\lambda(t)$, represents the average hazard function among all individuals (since the expectation of a randomly selected Z_i is equal to one). It follows that the subject-specific density and

survival functions are:

$$\begin{aligned} f(t|\mathbf{x}_i, Z_i) &= \lambda(t)\, Z_i \exp(\mathbf{x}_i'\boldsymbol{\beta})\, S(t|x_i, Z_i) \\ S(t|\mathbf{x}_i, Z_i) &= \exp\Big(- Z_i \exp(\mathbf{x}_i'\boldsymbol{\beta})\, \Lambda(t)\Big), \end{aligned}$$

where $\Lambda(t) = \int_0^t \lambda(t)\, dt$ denotes the cumulative baseline hazard at time t.

The interpretation of the hazard ratio, $\exp(\beta_p)$ for covariate x_p, in the proposed model differs from its interpretation in the standard Cox model. In the Cox model, the hazard ratio is interpreted as the average relative change in risk for a population given an increase in x_p of one unit. In the compound Poisson mixture frailty model, the hazard ratio measures the change in risk of an event for an at-risk individual (i.e. $Z_i > 0$). This interpretation remains the same if frailty is modeled with a single compound Poisson distribution (as in the model proposed in Chapter 2) instead of a mixture. The interpretation would also be similar if using a cure rate model with a frailty term for the non-cured fraction. By comparison, if modeling frailty without including a cured fraction (whether through a cure rate model or a frailty distribution like the compound Poission distribution that allows for no risk), the hazard ratio is interpreted as the effect on risk for any individual.

3.2.4 Model Likelihood

We next define the model likelihood. For each participant i, we observe t_i, the potentially censored event time and δ_i, an indicator variable equaling one if an event is observed and

zero otherwise. The likelihood for this model can be written as:

$$
\begin{aligned}
&L\big(\boldsymbol{\beta}, \lambda(t), \boldsymbol{\rho}, \boldsymbol{\eta}, \underline{\boldsymbol{\gamma}}, \boldsymbol{\alpha} | (t_i, \delta_i, \mathbf{x}_i, \mathbf{w}_i, R_i, N_i, Z_i), i = 1, \ldots, n\big) \\
&\quad = \prod_{i=1}^{n} f(R_i|\mathbf{w}_i, \underline{\boldsymbol{\gamma}}, \boldsymbol{\alpha})\, f(N_i|R_i, \boldsymbol{\rho})\, f(Z_i|R_i, N_i, \boldsymbol{\eta}, \nu)\, f(t_i, \delta_i|\mathbf{x}_i, Z_i, \boldsymbol{\beta}, \lambda(t)) \\
&\quad = \prod_{i=1}^{n} \Big(\prod_{k=1}^{K} \big(P(R_i = k|\mathbf{w}_i, \underline{\boldsymbol{\gamma}}, \boldsymbol{\alpha})\, f(N_i|R_i = k, \rho_k)\, f(Z_i|R_i = k, N_i, \eta_k, \nu)^{\mathbf{1}_{[N_i>0]}} \\
&\qquad\qquad f(Z_i|R_i = k, N_i, \eta_k, \nu)^{\mathbf{1}_{[N_i=0]}} \big)^{\mathbf{1}_{[R_i=k]}} \Big) \\
&\qquad \times f(t_i, \delta_i|\mathbf{x}_i, Z_i, \boldsymbol{\beta}, \lambda(t)))^{\mathbf{1}_{[N_i>0]}}\ f(t_i, \delta_i|\mathbf{x}_i, Z_i, \boldsymbol{\beta}, \lambda(t)))^{\mathbf{1}_{[N_i=0]}} \\
&\quad = \prod_{i=1}^{n} \Big(\prod_{k=1}^{K} \big(\big(G^{-1}(\alpha_k - \mathbf{w}_i\boldsymbol{\gamma}_k) - G^{-1}(\alpha_{k-1} - \mathbf{w}_i\boldsymbol{\gamma}_{k-1}) \big) \frac{\rho_k^{N_i}}{\Gamma(N_i+1)} e^{-\rho_k} \\
&\qquad\qquad \big(\frac{\nu^{N_i\eta_k}}{\Gamma(N_i\eta_k)} Z_i^{N_i\eta_k - 1} e^{-Z_i\nu} \big)^{\mathbf{1}_{[N_i>0]}} \times \boldsymbol{\Delta}_0(Z_i)^{\mathbf{1}_{[N_i=0]}} \big)^{\mathbf{1}_{[R_i=k]}} \Big) \\
&\qquad \times \big((\lambda(t) Z_i e^{\mathbf{x}_i'\boldsymbol{\beta}})^{\delta_i}\, e^{-Z_i e^{\mathbf{x}_i'\boldsymbol{\beta}}\, \Lambda(t_i)} \big)^{\mathbf{1}_{[N_i>0]}} \times \boldsymbol{\Delta}_\infty(t_i)^{\mathbf{1}_{[N_i=0]}}
\end{aligned}
\tag{3.3}
$$

where $\boldsymbol{\rho}$ is the vector of mean parameters $[\rho_1, \rho_2, \ldots, \rho_K]$, $\boldsymbol{\eta}$ is the vector of shape parameters $[\eta_1, \eta_2, \ldots, \eta_K]$, $\boldsymbol{\alpha}$ is the vector of intercepts $[\alpha_0, \alpha_1, \ldots, \alpha_K]$, $\underline{\boldsymbol{\gamma}}$ denotes parameter vectors $\boldsymbol{\gamma}_1, \ldots \boldsymbol{\gamma}_{K-1}$, and $\boldsymbol{\Delta}_0(Z_i)$, $\boldsymbol{\Delta}_\infty(t_i)$ indicate point mass for frailty, Z_i, at zero and time-to-event, t_i, at infinity, respectively, for participant i with $N_i = 0$ i.e. no risk. ν is the expectation of $\rho_k\eta_k$ for a randomly selected individual, as defined in Equation (3.2). While ν is a function of parameters $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$, and $\boldsymbol{\eta}$ and covariates $\mathbf{w}_1, \ldots, \mathbf{w}_n$, we have suppressed conditioning in the likelihood above for clarity in presentation.

Site-stratification of the baseline hazard function, $\lambda(t)$, is common for analysis of data from multi-center studies and can be accommodated by this model. If, however, site is included in the ordinal latent class regression model, it would be redundant to stratify estimation of the baseline hazard by site.

3.2.5 Bayesian Model Estimation

To complete Bayesian specification of the proposed model we specify prior distributions for modeling parameters before defining the posterior distribution and discussing sampling methods.

The application at hand will influence prior selection for $\boldsymbol{\rho}$, the mean parameter for the number of exposure processes in each latent class, as well as $\boldsymbol{\eta}$, the shape parameter for the level of exposure associated with each process. Uniform, gamma, and normal priors are all options for both parameters. In this presentation, we will use a gamma prior for each component of $\boldsymbol{\eta}$, $\boldsymbol{\pi}(\boldsymbol{\eta}) = Gamma(\mathbf{O}_{\boldsymbol{\eta}}, \mathbf{T}_{\boldsymbol{\eta}})$, as it naturally bounds the prior values for $\eta_1, \ldots, \eta_K$ above zero and its parameters can easily be specified to reflect prior knowledge or to be uninformative. Here, $Gamma(O, T)$ is the gamma distribution with shape O and rate T, $\mathbf{O}_{\boldsymbol{\eta}} = (O_{\eta_1}, \ldots, O_{\eta_K})$, and $\mathbf{T}_{\boldsymbol{\eta}} = (T_{\eta_1}, \ldots, T_{\eta_K})$.

For each ρ_k, rather than specifying priors on the mean number of exposure processes for each class k, we can instead place a prior on the proportion at risk, denoted here as $P_k(Risk)$ and defined as $P(N_i > 0 | R_i = k, \rho_k) = 1 - e^{-\rho_k}$. This approach can be used if we want to identify participants with a particular likelihood of exposure. For example, to identify participants with very high risk of an event, the prior distribution for the proportion at risk in the highest risk latent class should have greatest support closer to one. Setting a beta prior on the probability of risk in class k, $P_k(Risk) \sim Beta(A_k, B_k)$, where $Beta(A, B)$ is the beta distribution with shape parameters A and B, is equivalent to the following prior on ρ_k:

$$\boldsymbol{\pi}(\rho_k) = \frac{\Gamma(A_k + B_k)}{\Gamma(A_k)\Gamma(B_k)} \left(1 - \exp(-\rho_k)\right)^{A_k - 1} \exp(-\rho_k)^{B_k}.$$

Prior selection for other model parameters will be application-specific. Priors for the baseline hazard function, $\lambda(t)$, and proportional hazards regression model coefficients, $\boldsymbol{\beta}$, can be similar to priors typically used for survival models [27].

After specifying priors on all model parameters, we can identify the joint posterior distribution of the parameters and latent variables:

$$\begin{aligned} &\mathbf{p}\Big(\boldsymbol{\beta}, \lambda(t), \boldsymbol{\rho}, \boldsymbol{\eta}, \underline{\boldsymbol{\gamma}}, \boldsymbol{\alpha}, \mathbf{R}, \mathbf{N}, \mathbf{Z} \big| \big((t_i, \delta_i, \mathbf{x}_i, \mathbf{w}_i), i = 1, \ldots, n\big), \boldsymbol{\Theta}\Big) \\ &\propto L\big(\boldsymbol{\beta}, \lambda(t), \boldsymbol{\rho}, \boldsymbol{\eta}, \underline{\boldsymbol{\gamma}}, \boldsymbol{\alpha} \,|\, (t_i, \delta_i, \mathbf{x}_i, \mathbf{w}_i, R_i, N_i, Z_i), i = 1, \ldots, n\big) \times \boldsymbol{\pi}\big(\boldsymbol{\beta}, \lambda(t), \boldsymbol{\rho}, \boldsymbol{\eta}, \underline{\boldsymbol{\gamma}}, \boldsymbol{\alpha} \,|\, \boldsymbol{\Theta}\big) \end{aligned} \tag{3.4}$$

where $\mathbf{p}(\cdot|\text{data})$ denotes the joint posterior density, $\mathbf{R}$, $\mathbf{N}$, $\mathbf{Z}$ are vectors of latent variables R_i, N_i, Z_i, $i = 1, \ldots, n$, and $\boldsymbol{\pi}(\cdot|\boldsymbol{\Theta})$ denotes the joint prior density for the model parameters given hyperparameters $\boldsymbol{\Theta}$. The conditional posterior distributions for each parameter follow

from this full posterior; these are given Appendix C.1. A Gibbs sampling algorithm [21] is used to obtain samples from the model's posterior. Slice sampling [40] is used to update posterior samples when conjugate priors are not available. At each iteration of the sampling algorithm, latent variables R_i, N_i, and Z_i are also sampled for every participant $i = 1, \ldots, n$. Conditional posterior distributions for these latent variables are also detailed in Appendix C.1. Sampling is performed in `R` with independently written code.

3.3 Simulations

3.3.1 Methods

Simulations were done to assess model performance in a variety of situations. First, data were generated following the model proposed in Section 3.2 for a sample size of n =3,000 participants divided evenly between intervention and control arms. Then, model estimation was carried out for each particular scenario. Finally, posterior estimates for model parameters were compared to known values. Bayesian p-values for survival at six month intervals were also examined, as described in Section 2.2.3. The compound Poisson frailty model developed in Chapter 2 was also fit, in order to evaluate model performance when ignoring the latent class structure. We detail each step below.

Data Generation

Data generation occurred in three stages. First, participants were assigned to one of three risk classes, corresponding to low, medium, and high risk, following the latent class model presented in Section 3.2.1. For each simulation replication, latent classes were generated based on four binary risk-related covariates. Covariate values were correlated, i.e. if a participant had one risk factor, she was more likely to have another. Covariates were also constant across simulation replications. Classification was done under the proportional odds model, a version of the general cumulative link model given in Equation (3.1), where a logistic link is used and covariate effects are constant across classes, $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_k$ for all $k = 1, \ldots, K$. (The proportional odds model is described in more detail in Section 3.4.2.) Covariates $\mathbf{w}^* = [w_1^*, w_2^*, w_3^*, w_4^*]$ had the following effects on classification: $\exp(\boldsymbol{\gamma}^*) =$

$\left[\exp(\gamma_1^*), \exp(\gamma_2^*), \exp(\gamma_3^*), \exp(\gamma_4^*)\right]$ = [2, 1.5, 3, 0.5]. The intercept vector, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$, was fixed at [-0.5, 1] so that, given $\mathbf{w}^*$ and $\boldsymbol{\gamma}^*$, participants were evenly distributed across low, medium, and high risk classes. At each simulation replication, latent classes were sampled for each participant i from the multinomial distribution defined by covariates $\mathbf{w}_i^*$, effects $\boldsymbol{\gamma}^*$, and intercepts $\boldsymbol{\alpha}$, $i = 1, \ldots, n$. Covariate data and parameter values are given in detail in Appendix C.2.

Second, compound Poisson frailties were generated conditional on latent class membership, as described in Section 3.2.2. The values of ρ for each class corresponded to a 5%, 20%, and 80% probability of risk for an individual in the low, medium, and high risk classes, respectively. Values for η were fixed, in order, at 3, 4, and 5.

Third, event times were generated for each at-risk participant under the hazard model presented in Section 3.2.3. The event time for each participant with positive frailty was sampled from an exponential distribution with rate equal to the product of her frailty, Z_i, a constant baseline hazard, $\lambda = 5$ events per 100 person-years (PY), and a hazard ratio (HR) of 0.5 for those in the intervention arm, i.e. $\beta = \log(0.5)$. Censoring was imposed after 180 events (a 6% event rate). No other censoring was done.

Model Estimation

After data generation, model estimation was performed as described in Section 3.2.5 for the following scenarios:

(I) Correct priors: The prior mean was set equal to the true value for all model parameters other than the log-HR (which had a diffuse normal prior in all scenarios). Not all priors were centered around the true value, however, as scale parameters were set to allow for ample variability in sampling. Priors for the proportion at risk and η, for

instance, were necessarily skewed. The priors were as follows:

$$\text{Prior:} \begin{cases} \beta \sim N(0, 100) \\ \lambda_0 \sim Gamma(0.05 \times 50, 50) \\ 1 - \exp(-\rho_k) \sim Beta(\frac{P_k(Risk)}{1-P_k(Risk)} \times B_k,\, B_k) \\ \qquad\qquad \mathbf{P(Risk)} = [0.05, 0.25, 0.80],\, \mathbf{B} = [3,3,3],\, k = 1, \ldots, 3 \\ \eta_k \sim Gamma(O_{\eta_k}, T_{\eta_k}) \qquad \mathbf{O_\eta} = [4,6,8],\, \mathbf{T_\eta} = [2,2,2],\, k = 1, \ldots, 3 \\ \gamma_d^* \sim Normal(M_{\gamma_d^*},\, \Sigma^2_{\gamma_d^*}) \qquad \exp(M_{\boldsymbol{\gamma}^*}) = [2, 1.5, 3, 0.5], \\ \qquad\qquad \Sigma_{\gamma_d^*} = 0.5 \,\forall\, d,\, d = 1, \ldots, 4 \\ \alpha_k \sim Normal(M_{\alpha_k}, \Sigma^2_{\alpha_k}) \qquad M_{\boldsymbol{\alpha}} = [-0.5, 1],\, \Sigma_{\alpha_k} = 1 \,\forall\, k = 1, 2 \end{cases}$$

(II) Non-informative priors: Flat priors were placed on all model parameters. Priors were uniform on the scale used in (I) for all parameters, i.e. there was a $Uniform(0, 1)$ prior placed on the proportion at risk in each class k, $1 - \exp(-\rho_k)$, instead of a flat prior on the mean number of exposure processes, ρ_k. Flat priors on all model parameters were also specified for estimation of the compound Poisson frailty model.

(III) Leave out one predictor: Model estimation was performed with only three covariates. w_2, the risk factor associated with the highest risk ($\exp(\gamma_2^*) = 3$), was omitted. Priors for all other parameters were the same as in scenario (I).

(IV) Fit with two classes: The model was fit assuming two classes instead of three. The following priors were changed from scenario (I).

$$\text{Prior:} \begin{cases} 1 - \exp(-\rho_k) \sim Beta(\frac{P_k(Risk)}{1-P_k(Risk} \times B_k,\, B_k) \\ \qquad\qquad \mathbf{P(Risk)} = [0.15, 0.525],\, \mathbf{B} = [3,3],\, k = 1, 2 \\ \eta_k \sim Gamma(O_{\eta_k}, T_{\eta_k}) \qquad \mathbf{O_\eta} = [5,7],\, \mathbf{T_\eta} = [2,2],\, k = 1, 2 \\ \alpha \sim Normal(0, 1^2) \end{cases}$$

(V) Unbalanced classes: Data generation was changed so that class membership was not balanced. Intercept vector $\boldsymbol{\alpha}$ was set equal to [0.25, 1.25]. As a result, 51%,

21%, and 28% of participants were assigned to the low, medium, and high risk classes, respectively. Priors were specified as in scenario (I) except $M_{\boldsymbol{\alpha}} = [0.25, 1.25]$.

(VI) Single compound Poisson frailty: The latent class model was estimated with covariates $\mathbf{w}^*$ and three latent risk classes for data generated with a single compound Poisson frailty distribution. The data generating mechanism described in Section 2.3.1 was used with 25% of participants at risk and a 6% event rate. Priors for scenario (I) were used. Posterior estimates for the HR and baseline hazard were compared to those reported in Section 2.3.2.

The compound Poisson frailty model developed in Chapter 2 was also applied to data simulated as described in Section 3.3.1. This model was estimated under scenarios (I), (II), and (V) and compared to estimates from the latent class compound Poisson mixture frailty model. For scenarios (I) and (V), the priors for the proportion at risk and η were based on the average proportion at risk and η across all classes. The priors specified for scenario (I) were:

$$\text{Prior:} \begin{cases} \beta \sim N(0, 100) \\ \lambda_0 \sim Gamma(0.05 \times 50, 50) \\ 1 - \exp(-\rho) \sim Beta(\frac{\bar{P}(Risk)}{1-\bar{P}(Risk)} \times 3,\ 3) \qquad \bar{P}(Risk) = 36\% \\ \eta \sim Gamma(6, 2) \end{cases}$$

For scenario (II), flat priors were specified for all model parameters.

Posterior medians and 95% quantile-based credible intervals were identified for all model parameters. Bayesian p-values comparing survival at six month intervals were also calculated. Each simulation was repeated 500 times. Posterior samples were the result of 25,000 iterations of which the first 10,000 were discarded for burn-in.

3.3.2 Results

Tables 3.1 and 3.2 show simulation study results. Table 3.1 gives estimates and interval coverage for the HR, $\exp(\beta)$, and the baseline hazard, λ, for both the latent class (LC) compound Poisson mixture frailty model estimation and estimation with a single compound

Scenario	Model	HR Est	HR Cov	λ Est	λ Cov	Bayesian p-values 6mo	12 mo	18mo
(I): Correct priors	LC	0.50	95%	5.0	97%	0.51	0.49	0.48
	CP	0.50	95%	5.0	97%	0.50	0.49	0.48
(II): Non-informative priors	LC	0.50	95%	5.0	92%	0.50	0.49	0.50
	CP	0.50	95%	5.0	90%	0.50	0.50	0.50
(III): Leave out 1 predictor	LC	0.50	95%	5.0	97%	0.52	0.49	0.47
(IV): Fit with 2 classes	LC	0.51	95%	4.8	93%	0.44	0.48	0.52
(V): Unbalanced groups	LC	0.50	94%	5.0	97	0.50	0.49	0.50
	CP	0.50	95%	5.0	90%	0.51	0.49	0.49
(VI): Single CP Frailty	LC	0.52	95%	4.7	92%	0.43	0.47	0.55
	CP	0.50	96%	5.0	97%	0.53	0.48	0.48
Truth		0.50		5.0 per 100 PY				

Table 3.1: Average posterior median for HR and baseline hazard λ and coverage of 95% credible intervals from estimation with the proposed latent class CP mixture frailty model (LC) and single CP frailty model (CP) for various simulation scenarios. Bayesian p-values for survival at six month intervals also given. Results are from 500 replications.

Poisson (CP) frailty model. Both models give unbiased estimates and nominal coverage of the HR and baseline hazard with correct (I) and non-informative (II) priors when participants are evenly distributed across latent classes. We also see unbiased estimation and nominal coverage of these parameters from both models when groups are less balanced (V). Furthermore, estimation and coverage of the HR and baseline hazard appear robust to model misspecification, as seen in scenarios (III) and (IV). Finally, the HR estimate from the latent class model when data are generated with a single compound Poisson frailty distribution is only slightly biased while coverage of 95% credible intervals is still nominal.

Bayesian p-values for survival at six-month intervals are also given in Table 3.1 for both estimation approaches in all scenarios. All hover around 0.50, indicating that survival under the posterior predictive distribution in all cases is similar to the observed (simulated)

survival.

Table 3.2 gives estimates and interval coverage for the proportion at risk and η for each class in all scenarios. We see that the estimated proportion at risk and η for each class are similar to the true values when correct priors are used (I), as well as when one predictor of the latent class model is omitted, as in scenario (III). We note that the proportion at risk and η are both underestimated in the low risk class for both scenarios (I) and (III) and that coverage of 95% credible intervals is low for the proportion at risk. This bias is likely due in part to the skewed priors placed on those parameters. Classification probabilities are also similar to true classification for scenarios (I) and (III). Similar results were seen when data were generated with unbalanced groups (scenario (V)). Posterior classification probabilities in scenario (V) were not unbiased estimates of true classification probabilities, however, as the posterior erred toward a more even distribution across class.

Alternatively, when non-informative priors are used, as in scenario (II), estimation of class-specific frailty parameters for all classes are biased and 95% credible intervals show poor coverage. In particular, the estimates for η are quite large, a problem we also encountered when using non-informative priors on the single CP frailty model in Section 2.3.2. Posterior classification also shows uneven distribution of participants in this scenario. In the absence of any prior information, more participants were placed in the middle risk class and fewer were placed in the higher risk class.

In scenario (IV), the latent class model was fit with two classes instead three (the true number of classes), we see that nearly all the observations were classified into one group. In practice, when the number of classes is unknown, such unbalanced classification would be an indication that a different number of classes should be specified or that the single CP frailty model may be more appropriate. In this specific study, one may reasonably choose to assume a single class after observing that the model with two latent classes placed nearly all participants into one group. There was no indication that the data were generated using three latent classes and, subsequently, that a model assuming three latent classes was correct. Similarly, when we fit the latent class model with three classes to data that was generated with a single CP frailty distribution (scenario (VI)), there is no indication that the model used for estimation is incorrect, as all parameter estimates and class groupings

Scenario	Latent Class	$P(Risk)$ Est	$P(Risk)$ Cov	η Est	η Cov	Class. Prob. Est	Class. Prob. Cov
	Low	0.0035	73%	1.5	100%	0.31	99%
(I): Correct priors	Med	0.26	100%	2.8	100%	0.35	100%
	High	0.81	96%	4.4	100%	0.33	95%
	Low	0.33	81%	6.4	78%	0.35	79%
(II): Non-informative priors	Med	0.61	53%	15	26%	0.47	81%
	High	0.82	90%	29	4.8%	0.16	47%
	Low	0.0042	75%	1.5	100%	0.33	100%
(III): Leave out 1 predictor	Med	0.31	100%	2.8	100%	0.35	100%
	High	0.81	100%	4.3	100%	0.31	98%
	Low	0.071	n/a	2.0	n/a	0.0020	n/a
(IV): Fit 2 classes	High	0.52	n/a	3.7	n/a	>0.99	n/a
Truth		(0.05, 0.25, 0.80)		(2,3,4)		(0.35, 0.33, 0.32)	
	Low	0.0058	76%	1.5	100%	0.44	96%
(V): Unbalanced groups	Med	0.29	100%	2.8	100%	0.28	100%
	High	0.81	100%	4.4	100%	0.27	99%
Truth		(0.05, 0.25, 0.80)		(2,3,4)		(0.51, 0.21, 0.28)	
	Low	0.15	n/a	0.70	n/a	0.34	n/a
(VI): Single CP Frailty	Med	0.50	n/a	1.7	n/a	0.40	n/a
	High	0.78	n/a	3.1	n/a	0.24	n/a
Truth		0.25		2		n/a	

Table 3.2: Average posterior median and coverage of 95% credible intervals for the proportion at risk and η in each latent risk class. The average posterior medians and interval coverage for the proportion of participants assigned to each class are also given. Results are from 500 simulation replications.

seem reasonable. If the true data generating mechanism were unknown, the only suggestion we would have that the CP frailty model was more appropriate than the LC model would be that the Bayesian p-values for the CP model (given in Table 3.1) were closer to 0.50 than those for the LC model at every time interval considered. However, since test quantities for Bayesian p-values are arbitrarily chosen, it would be unwise to depend soley on these measures for model selection. Thus, this simulation study gives us limited guidance in selecting the number of classes.

In this set of simulation studies, we see that model estimation assuming a single CP frailty distribution provides an unbiased estimate of the hazard ratio and 95% credible intervals with nominal coverage, even when data are generated with a mixture of compound Poisson frailty distributions. The latent class approach, under this particular data-generating scheme, does not provide any advantage in the estimation of the hazard ratio or baseline hazard function. In fact, it appears that the latent class approach's chief advantage is that it also identifies latent risk classes. Thus, the compound Poisson mixture frailty model being presented in this chapter is recommended for modeling heterogeneity in situations where identifying latent risk classes is of particular interest and appropriate informative priors are available. To the latter point, placing moderately informative priors on the majority of model parameters is easily justifiable in many situations. Typically, prior information about whether covariates increase or decrease risk is readily available. Furthermore, placing priors on class-specific frailty model parameters is actually a strength of the proposed model as it informs classification. In particular, specifying more informative priors on the proportion at risk in each class enables us to identify groups of participants with relatively low or high risk of an event.

3.4 Application

We illustrate the proposed model with analysis of data from two HIV prevention trials: Partners PrEP and VOICE.

3.4.1 Data

The Partners PrEP study was a phase III trial of oral antiretroviral (ARV) treatment as pre-exposure prophylaxis for HIV-1 transmission in serodiscordant couples [8]. From July 2008 through November 2010, the study enrolled 4,758 heterosexual couples in which one partner was HIV-infected and the other partner was not from nine sites in Kenya and Uganda. Seronegative partners were randomized to one of three study arms: oral tenofovir (TDF), combination tenofovir and emtricitabine (TDF-FTC, also known as Truvada®), or placebo, all once-daily, and were tested monthly for seroconversion for up to thirty-six months. The study's primary analysis estimated the effectiveness (1-HR) of each intervention in reducing risk of seroconversion compared to placebo. On July 12, 2011, the trial's placebo arm was discontinued because both active drugs demonstrated protection that exceeded predetermined stopping boundaries. Using data collected through this stopping date, TDF and TDF-FTC were estimated to reduce risk of HIV acquisition by 67% (95% CI: 44, 81%) and 75% (95% CI: 55, 87%) , respectively.

The VOICE (Vaginal and Oral Interventions to Control the Epidemic) study was a phase IIb trial of the safety and effectiveness of two ARV-based approaches for preventing heterosexual transmission of HIV in women: oral ARV tablets or vaginal microbicide containing ARV in gel form [33]. Beginning in September 2009, 5,029 sexually active women were enrolled in the study from fifteen sites in and around Durban and Johannesburg, South Africa, Kampala, Uganda, and Harare, Zimbabwe. Participants were randomized into one of five study arms: oral TDF, TDF-FTC, oral placebo, tenofovir-based gel, and placebo gel, all once-daily, and were tested quarterly for HIV seroconversion. The study's primary analysis estimated the effectiveness of each active ARV arm in comparison to the appropriate placebo arm. The oral TDF arm was suspended on September 16, 2011, and the TDF and placebo gel arms were stopped on November 17, 2011, both due to futility. Risk of seroconversion was reduced among participants in the TDF-FTC arm, but this result was not significant.

3.4.2 Model and Prior Specification

As discussed in Chapter 2, the compound Poisson distribution reflects the risk mechanism for HIV seroconversion. In these analyses, we use the proposed latent class approach to model frailty with a compound Poisson mixture distribution in order to estimate the individual-level effect of each intervention (oral TDF and TDF-FTC for Partners PrEP and oral TDF, TDF-FTC, and gel TDF for VOICE) on risk of seroconversion. Survival for participants randomized to oral interventions (TDF and TDF-FTC) was compared to survival under oral placebo for both trials. In the VOICE analysis, the TDF gel arm was compared to the placebo gel arm. A single model was used for all comparisons in the VOICE trial in order to share information across participants about site and covariate effects. In both analyses, $x_{pi} = 1$ indicated randomization to study arm p for individual i, $x_{pi} = 0$ indicated otherwise, and the HR associated with each oral intervention p was $\exp(\beta_p)$ while the HR associated with TDF gel in VOICE was $\exp(\beta_{\text{TDF-gel}} - \beta_{\text{placebo-gel}})$. Diffuse normal priors were specified for the log-HR of each intervention.

As was specified in the primary analysis of these studies, participants who were infected at baseline but did not seroconvert until after starting the study were excluded from the analysis. For participants without seroconversion observed, time until the date of last negative test prior to the suspension of a trial arm was taken as the censored event time. For participants with seroconversion observed before the first trial arm was suspended in each study, event time was sampled at each iteration of the Gibbs sampler using slice sampling techniques and a uniform prior on the interval covering time between the participant's last negative and first positive serology test.

Due to the low number of events in the Partners PrEP trial, we classified participants into risk groups with a *proportional odds* model, a cumulative logistic-link model that assumes constant covariate effects across classes. The proportional odds model is more efficient than the general cumulative link model presented in Section 3.2.1, and the mean squared error may be smaller, even in some cases where the proportional odds assumption is violated if the difference in estimated effects is small in practical terms [7]. The proportional odds

model can be expressed as

$$\text{logit}\big(P(R_i \leq k|\mathbf{w^*}_i)\big) = \alpha_k - \mathbf{w^*}_i'\boldsymbol{\gamma}^* \tag{3.5}$$

where $\boldsymbol{\gamma}^*$ is a D^*-length vector of effects associated with covariates $\mathbf{w^*}_i$ that is constant across all k, $k = 1, \ldots (K-1)$. Comparing the proportional odds model in Equation (3.5) to the generalized cumulative link model in Equation (3.1) with G^{-1} as the logit link, we note that if the effects in parameter vector $\boldsymbol{\gamma}_k$ were equal across all k then the parameter vector we've denoted $\boldsymbol{\gamma}*$ would also be equal to all $\boldsymbol{\gamma}_k$. We use the $(^*)$ notation to distinguish between covariates for which proportionality is assumed, as in Equation (3.5), and when it is not, as in the more general model given in Equation (3.1).

In a general population study like VOICE, we expect the probability and magnitude of risk to vary by site based on factors like the area's seroprevalence, behavioral norms, and other local characteristics. In fact, seroincidence in VOICE did differ considerably across site locations. For these reasons, we chose to include site location in the risk classification model for the VOICE analysis. Additionally, we chose to forgo the proportional odds assumption for location effect, resulting in a *partial proportional odds* model. The partial proportional odds model assumes a constant effect across class for some but not all covariates [7]:

$$\text{logit}\big(P(R_i \leq k|\mathbf{w}_i, \mathbf{w^*}_i)\big) = \alpha_k - (\mathbf{w}_i'\boldsymbol{\gamma}_k + \mathbf{w^*}_i'\boldsymbol{\gamma}^*)$$

where $\boldsymbol{\gamma}_k$ is a vector of effects associated with covariates $\mathbf{w}_i$ that varies across k, $k = 1, \ldots, K-1$. For the VOICE analysis, $\mathbf{w}_i$ was a vector indicating site location for each participant and proportional odds were assumed for all other risk-related covariates, $\mathbf{w^*}_i$.

Covariates that (1) have been identified in the literature as risk factors for HIV infection and (2) were significant predictors of seroconversion in a univariate logistic regression model were included in the latent class model as covariates $\mathbf{w}^*$ for each analysis [28]. For the VOICE analysis, these covariates included age, parity, marital status, gonorrhea infection at baseline, and whether the participant's primary sexual partner had other partners. In the Partners PrEP analysis, covariates included gender, age, and number of children of the seronegative partner and the HIV-infected partner's CD4 count at baseline. Notably, study site was not a significant predictor of seroincidence among Partners PrEP participants. This

lack of association between site and seroconversion in the Partners PrEP trial is mainly because, in contrast to the VOICE trial, serostatus of a participant's primary partner is known and does not vary across sites.

For each proportional covariate effect $\gamma_1^*, \ldots, \gamma_{D*}^*$ and intercept $\alpha_1, \ldots, \alpha_{K-1}$, a normal prior with mean zero was specified in the latent class models for both analyses. Normal priors centered at zero were also used for the non-proportional effects of location on classification, $\boldsymbol{\gamma}_{kd}$, in the VOICE analysis, for $k = 1, \ldots, K-1$ and site location d, $d = 1, \ldots, D$.

For analysis of data from both trials, participants were classified into three risk groups reflecting low, medium, and high risk levels. Three classes were chosen in order to minimize the number of parameters to be estimated while obtaining a more detailed classification of participants than would be provided by two risk groups. In particular, the aim was to identify participants with particularly high and low risk. The expected distribution of participants across classes is influenced by the observed covariate values as well as the priors on cumulative link model parameters ($\boldsymbol{\alpha}$, $\boldsymbol{\gamma}^*$, and, when applicable, $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{K-1}$). In both models, the prior expected proportion of participants in each class was relatively even, due to normal priors with mean zero on cumulative link model parameters. However, classification in the low or high risk class was slightly more likely than in the medium class, owing to the distribution of covariate values among study participants.

Priors on the compound Poisson frailty model parameters for each latent class were specified in order to identify three distinct distributions for risk. Specifically, the prior distribution on the proportion at risk in the lowest risk class had highest probability near zero, the prior distribution for the medium risk class was centered around 50%, and the prior distribution for the highest risk class had highest probability close to 100%. Similarly, the prior distributions for the level of risk associated with each exposure process were stochastically ordered. The prior densities for the proportion at risk and η for each class are shown in Figure 3.1. Following the discussion in Section 2.2.1, beta priors were specified for the proportion at risk in each class, $1 - \exp(-\rho_k)$, and gamma priors were specified for each η_k, $k = 1, 2, 3$.

We allowed for flexible estimation of the baseline hazard function in both analyses by

assuming a piecewise constant function for $\lambda(t)$:

$$\lambda(t) = \lambda_s, \qquad t_s \leq t < t_{s+1},\ s = 1, \ldots, S \tag{3.6}$$

where t_s, $s = 1, \ldots, S$, denote the cut points, $t_1 = 0$, and $t_{S+1} = \infty$. Cut-points were determined by events with roughly eight events per interval in the Partners PrEP analysis and ten events per interval in the VOICE analysis. Since site location was included in the risk classification model for VOICE, estimation of the baseline hazard function was not stratified by site because, as previously mentioned, to do so would be redundant. In the Partners PrEP study, we did not expect site to be an important predictor of exposure or risk levels for participants and, so, did not use site for classification. While site-stratification of survival models is the standard for multisite clinical trials and was used in the primary analysis of this study, we chose not to stratify baseline hazard estimation by site because there were too few events in many of the sites. (Three of the nine sites had five or fewer events.) For both analyses, conjugate gamma priors were specified for each λ_s, $s = 1, \ldots, S$.

Prior distributions for all parameters in both models were specified in such a way that the expected incidence in the placebo arm under the prior was consistent with the incidence observed in control arms of recent, similarly conducted HIV prevention trials. Figure 3.2 shows the prior distribution of seroincidence in the placebo arm for the Partners PrEP and VOICE analyses, produced by 500 replications of a simulation generating event times given the specified priors and study participants' covariates under the latent class (LC) compound Poisson mixture model. Seroincidence observed in placebo arms of similar trials are indicated with plotting symbols. All trials included for comparison were population-based studies of ARV-based interventions, except Celum et al. 2008, a serodiscordant partners study of the effect of HSV-2 treatment on seroconversion, which we include to provide a comparison to the Partners PrEP study. The lower seroincidence under the prior for the Partners PrEP trial was intentional because we expect to observe a lower incidence in serodiscordant partner studies than general population studies. As shown in the figure, Celum et al. 2008 observed a seroincidence of 3.3 events per 100 person-years in the placebo arm. We see that the distribution of expected prior incidence in the placebo arm for the VOICE analysis is consistent with that which has been observed in previous, similar trials.

The priors on model parameters were specified as follows for the Partners PrEP analysis:

$$\text{Prior:} \begin{cases} \beta_p \sim N(0, 10^2) & \text{for } p = 1, 2 \\ \lambda_s \sim Gamma(0.04 \times 5, 5) & \text{for } s = 1, \ldots, S = 11 \\ 1 - \exp(-\rho_k) \sim Beta(A_k, B_k) & \mathbf{A} = [1, 4, 6],\ \mathbf{B} = [4, 4, 2],\ k = 1, \ldots, 3 \\ \eta_k \sim Gamma(O_{\eta_k}, T_{\eta_k}) & \mathbf{O_\eta} = [3, 4, 5],\ \mathbf{T_\eta} = [1, 1, 1],\ k = 1, \ldots, 3 \\ \alpha_k \sim Normal(0, 1^2) & \text{for } k = 1, 2. \\ \gamma_d^* \sim Normal(0, 0.75^2) & \text{for } d = 1, \ldots, D^* = 6 \end{cases}$$

The priors on model parameters for the VOICE analysis were:

$$\text{Prior:} \begin{cases} \beta \sim N(0, 10^2) & \text{for } p = 1, \ldots, 4 \\ \lambda_s \sim Gamma(0.0575 \times 10, 10) & \text{for } s = 1, \ldots, S = 21 \\ 1 - \exp(-\rho_k) \sim Beta(A_k, B_k) & \mathbf{A} = [1, 4, 6],\ \mathbf{B} = [4, 4, 2]\, k = 1, \ldots, 3 \\ \eta_k \sim Gamma(O_{\eta_k}, T_{\eta_k}) & \mathbf{O_\eta} = [3, 4, 5],\ \mathbf{T_\eta} = [1, 1, 1],\ k = 1, \ldots, 3 \\ \alpha_k \sim Normal(0, 1^2) & \text{for } k = 1, 2 \\ \gamma_d^* \sim Normal(0, 0.75^2) & \text{for } d = 1, \ldots, D^* = 7 \\ \gamma_{kd} \sim Normal(0, 1^2) & \text{for } k = 1, 2,\ d = 1, \ldots, D = 3. \end{cases}$$

To obtain posterior samples from the compound Poisson mixture frailty model, we repeated the Gibbs sampling algorithm (described in Section 3.2.5) for 100,000 iterations. The first 20,000 samples were discarded as burn-in, and the remaining samples from each parameter's posterior distribution were thinned, keeping every tenth sample. Parallel chains were run from different starting values to confirm that models converged to similar estimates. Model diagnostics performed included examining trace and cumulative distribution plots. Additionally, Bayesian p-values comparing survival under the posterior predictive distribution to that which was observed were calculated, as outlined in 2.2.3.

The posterior median was identified for each component of $\boldsymbol{\beta}$, as well as the surrounding 95% highest posterior density (HPD) intervals. These estimates were compared to point estimates and 95% model-based confidence intervals from a stratified Cox model. Posterior classification probabilities for risk-related covariate values were also summarized, as well as the posterior incidence, proportion at risk, and mean frailty across classes, using sampled

values of latent variables (R_i, N_i, Z_i), $i = 1, \ldots, n$ at every saved iteration of the Gibb's sampler.

3.4.3 Stratified Compound Poisson Frailty Model

Data from the Partners PrEP and VOICE trials were also analyzed using the stratified compound Poisson (CP) frailty model proposed in Chapter 2.

The Partners PrEP analysis was stratified by country (Kenya and Uganda) because there were too few events in some sites to stratify the analysis by individual sites. A piecewise constant baseline hazard function, as in Equation (3.6), was used for each country, and intervals for the piecewise function were shared across strata. Correlated priors on the baseline hazard function, as presented in Section 2.4.1 were placed on all λ_{sj}, the constant baseline hazard for country j at time t between t_s and t_{s+1}.

The VOICE analysis was stratified by site location. As in Chapter 2, we also used a beta prior on the proportion at risk at each location, with mean equal to the screening prevalence at that location. Screening prevalence at each site location for VOICE is given in Table 3.3. Unlike the Partners PrEP analysis in this chapter, we did not use correlated priors for location-specific piecewise constant baseline hazard functions. Since over 90% of seroconversions occurred at sites in and around Durban, South Africa, event times from those sites would dictate intervals for the piecewise function, resulting in many intervals for non-Durban sites with no events. So, independent priors were used for each location's baseline hazard function. We specified a piecewise constant baseline hazard function for Durban sites and a constant baseline hazard function for the other site locations, as there were not enough events in these locations to estimate the baseline hazard for multiple intervals. The number and percent of participants enrolled in each site with seroconversion observed before censoring are given Table 3.3.

Prior specification for all model parameters was again performed with the intention that expected seroincidence in the placebo arm be consistent with seroincidence observed in the placebo arms of previous, similar trials. The expected distribution of seroincidence in the placebo arm under the priors specified for the Partners PrEP and VOICE analyses are

Site location	Prevalence (%)	Sample size	# Events (%)
Durban, South Africa	25	3,329	192 (2.7%)
Johannesburg, South Africa	12	670	10 (1.5%)
Kampala, Uganda	15	320	6 (1.9%)
Harare, Zimbabwe	14	626	3 (0.48%)

Table 3.3: HIV prevalence at screening, sample size, and number of events (% of enrolled individuals with events) for site locations in the VOICE trial.

shown in Figure 3.2. Prior seroincidence for the site-stratified CP frailty models are given in gray.

The priors on model parameters were specified as follows for the Partners PrEP analysis:

$$\text{Prior:} \begin{cases} \beta \sim N(0, 100) \\ \lambda_{sj} | \rho_j, \theta_s \sim Gamma(\theta_s(1 - \exp(\rho_j)) \times \tau,\, \tau) \\ \qquad\qquad\qquad\qquad \text{for countries } j\text{=1, 2, intervals } s = 1, \ldots, S \\ 1 - \exp(-\rho_j) \sim Beta(0.8, 2) \qquad \text{for } j\text{=1,2} \\ \eta_j \sim Gamma(2, 0.5) \qquad \text{for } j\text{=1,2} \\ \theta_s \sim Gamma(5, 40) \qquad \text{for intervals } s = 1, \ldots, S \\ \tau \sim InverseGamma(2, 20) \end{cases}$$

and the VOICE analysis:

$$\text{Prior:} \begin{cases} \beta \sim N(0, 100) \\ \lambda_{sj} \sim Gamma(0.075 \times 20,\, 20) \\ \qquad\qquad\qquad \text{for locations } j = 1, \ldots, 4, \text{ intervals } s = 1, \ldots, S \\ 1 - \exp(-\rho_j) \sim Beta(2 \times \frac{Prev_j}{1 - Prev_j}, 2) \qquad \text{for } j = 1, \ldots, 4 \\ \eta_j \sim Gamma(2, 0.5) \qquad \text{for } j = 1, \ldots, 4 \end{cases}$$

where $Prev_j$ is the screening prevalence at each site location, given in Table 3.3.

3.4.4 Results

We applied the proposed model to data from two trials of ARV-based interventions for the prevention of HIV with two aims in mind: to obtain estimates of each intervention's effectiveness for at-risk individuals that are adjusted for heterogeneity and to examine how participants were categorized based on observed covariates. We first discuss effectiveness estimates before examining results from the latent risk classification.

Intervention Effectiveness

Table 3.4 reports population- and individual-level estimates of the effectiveness of oral TDF-FTC and TDF in the Partners PrEP trial. The estimated effectiveness of both interventions in reducing risk of seroconversion for at-risk individuals obtained from the proposed latent class-based compound Poisson (CP) mixture frailty model (77% and 70% for TDF-FTC and TDF, respectively) were slightly higher than the population-averaged effectiveness estimates from the Cox model (75% and 67%) as well as individual-level estimates from the country-stratified CP frailty model (75% and 66%). 95% HPD intervals around both estimates from the proposed model are slightly higher than model-based 95% CIs from the Cox model and narrower than 95% HPD intervals from the country-stratified CP frailty model.

Model	Oral TDF-FTC	Oral TDF
Country-stratified Cox	0.75 (0.55, 0.87)	0.67 (0.44, 0.81)
Country-stratified CP Frailty	0.75 (0.55, 0.90)	0.66 (0.40, 0.84)
Latent Class CP Mix. Frailty	0.77 (0.58, 0.88)	0.70 (0.50, 0.83)

Table 3.4: Estimated effectiveness of ARV-based interventions in Partners PrEP trial.

It was unexpected that the difference between estimates from the proposed model and Cox model was slight. When heterogeneity in risk is present, the individual-level effectiveness of a beneficial intervention is higher than the effectiveness averaged over the population, and we expected estimates from the proposed model to reflect this. Observing too

few events relative to the degree of heterogeneity present could explain why the country-stratified CP frailty model estimate of individual-level effectiveness was not higher than the population-level estimate, especially since the risk may not have varied much across countries. Furthermore, the Partners PrEP trial was stopped early so the targeted number of events was not reached. We recall from the simulation studies presented in Chapter 2 that the difference between Cox model population-averaged estimates and individual-level estimates of effectiveness from the compound Poisson frailty model are more similar at lower event rates. Yet, for the proposed latent class mixture frailty model, we expected that using risk-related covariate information to inform frailty modeling would have a greater impact on estimates of individual-level effectiveness.

Effectiveness estimates for each ARV-based intervention in VOICE are given in Table 3.5. The estimated effectiveness of TDF-FTC for an individual at risk is similar across all three models (between 12 and 15%). As with the Partners PrEP analysis, we had expected the estimate to be considerably larger for the frailty model than for the Cox model. Effectiveness estimates for TDF gel also did not vary much across models (between 3.3 and 4.3%). Furthermore, given poor adherence in the VOICE trial, we suspected that the negative result for oral TDF was the result of poor adherence rather than actual harm related to the drug and had anticipated that an estimate adjusted for frailty would show a smaller negative effect [34]. Yet, both the proposed latent class CP mixture frailty model and the location-stratified CP frailty model estimated oral TDF to be more harmful on the individual level than the population level. It is possible that risk-related covariates included in the latent class model were also associated with adherence such that participants who were likely to adhere would also be more likely to be classified in lower risk groups and assigned no risk or lower frailties. Underestimating risk for adherent individuals would have, subsequently, underestimated the intervention's effectiveness for those individuals. This association between adherence and covariates used in risk classification could have affected HR estimates for all trial interventions, though it was the most obvious in the oral TDF arm. Also, a similar effect would be expected in the stratified CP frailty analysis if drug adherence was strongly associated with study site location.

Model	Oral TDF-FTC	Oral TDF	TDF Gel
Location-stratified Cox	0.14 (-0.40, 0.47)	-0.48 (-1.3, 0.037)	0.043 (-0.43, 0.36)
Location-stratified CP Frailty	0.15 (-0.38, 0.46)	-0.59 (-1.4, -0.036)	0.033 (-0.49, 0.43)
Latent Class CP Mix. Frailty	0.12 (-0.41, 0.44)	-0.67 (-1.5, -0.14)	0.042 (-0.43, 0.44)

Table 3.5: Estimated Effectiveness (1-HR) of ARV-based interventions in VOICE trial.

Latent Risk Classification

In addition to estimating individual-level effectiveness, we wanted to assess how well the latent class model was able to distinguish between low, medium, and high risk groups. To do that, we compared the posterior distribution of seroincidence across groups as well as the average frailty and proportion at risk in each group. Table 3.6 gives posterior medians (95% credible intervals) for these measures, as well as overall classification frequencies, obtained from posterior samples of $\mathbf{R}$. This information is also displayed visually in Figures 3.3 and 3.4. We see that, in analysis of data from both trials, the model was successful in identifying separate risk groups as the incidence, proportion at risk, and frailty vary considerably between classes. For example, in the analysis of the VOICE data, the median posterior incidence for participants in the lowest risk group was zero while the median posterior incidence for participants in the highest risk group was 10 events per 100 person-years (PY). The percentages of participants in each group with some risk (i.e. frailty above zero) were also consistent with low, medium, and high risk groups.

For both analyses, we were surprised to see that the vast majority of participants were classified into the medium and high risk classes. Since we believe that most participants had no or very low risk during these trials, we expected to see more participants classified into the low risk group. The distribution of covariates across participants, however, provides some explanation for the posterior classification distributions. Median posterior classification frequencies for risk-related covariates, obtained from posterior samples of $\mathbf{R}$, are given in Tables 3.7 and 3.8. These results are also displayed visually in Figure 3.5 for Partners PrEP

and Figures 3.6- 3.8 for VOICE. For the VOICE analysis, we see that site location and marital/cohabitation status are likely driving overall classification frequencies. Participants from study sites in and around Durban, South Africa had a high probability of being in the medium or high risk classes because over 90% of the observed events were among these participants. Since two-thirds of all study participants were in Durban sites, average posterior classification probabilities also favor medium and high risk classes. A similar effect is observed for participants who are unmarried and not living with a partner, since two-thirds of study participants were unmarried or not cohabitating and nearly 90% of events were observed in that group. In the Partners PrEP analysis (Table 3.7), no subset of the covariates seems to be driving classification and the median posterior probability of classification in the high risk group was 56%. Even among low risk covariates like older age, male sex, and more children, the probability of classification in the high risk group was nearly half. Yet, since the proposed model allows for no risk even in the highest risk groups, we see that exposure in the high risk group is not certain as the median posterior probability of any risk in the high risk group is only 57%.

	Latent Risk Classes		
	Low	Medium	High
Partners PrEP			
Incidence (per 100 PY)	0 (0,0.24)	0.26 (0,1.1)	1.3 (1.7, 3.1)
Percentage at Risk	11% (0.59, 39%)	38% (17, 60%)	57% (44, 73%)
Average Frailty	0.029 (0.0010, 0.18)	0.25 (0.062, 0.95)	1.6 (1.2, 2.8)
Overall Classification Frequency	17% (6.3, 37%)	25% (4.6, 61%)	56% (26, 77%)
VOICE			
Incidence (per 100PY)	0 (0, 0.73)	0.76 (0.22, 1.9)	10 (15,23)
Percentage at Risk	15% (0.76, 45%)	21% (47, 69%)	69% (50, 87%)
Average Frailty	0.021 (0.00063, 0.094)	0.13 (0.049, 0.32)	2.4 (1.6, 3.7)
Overall Classification Frequency	13% (2.1, 31%)	48% (22, 67%)	38% (25, 59%)

Table 3.6: Posterior medians (95% credible interval) of summary measures for latent risk classes.

	n	Incidence	Classification Probabilities		
Covariate	(%)	(per 100PY)	Low	Medium	High
All	4,707 (100%)	1.0	17%	25%	56%
Age (years)					
≤ 20	84 (1.8%)	5.1	7.1%	14%	75%
21-30	1,620 (34%)	1.5	8.6%	18%	72%
> 30	3,003 (64%)	0.73	22%	29%	47%
Sex					
Female	1,776 (38%)	1.5	8.7%	20%	70%
Male	2,931 (63%)	0.75	22%	29%	47%
# **Children**					
None	396 (8.4%)	1.6	12%	21%	65%
1 or 2	1,442 (31%)	1.4	10%	18%	70%
≥ 3	2,869 (61%)	0.83	22%	29%	48%
CD4 count (cells/mm^3)					
Mean (SE)	555 (3.6)		729 (88)	581 (26)	491 (26)

Table 3.7: Median posterior classification probabilities across covariate values for **Partners PrEP** trial.

Model Assessment

Model fit was assessed by comparing predictive posterior survival to observed survival. Figures 3.9 and 3.11 show Kaplan-Meier curves (bold lines) for observed survival data along with survival data generated under the posterior predictive distribution of the latent class compound Poisson mixture frailty model (gray lines) for the Partners PrEP and VOICE analyses, respectively. Figures 3.10 and 3.12 show observed and predicted survival under the stratified compound Poisson frailty proposed in Chapter 2 for analyses of the two trials.

Bayesian p-values comparing observed to predicted survival at six-month intervals are also given. In analysis of both trials, we see more variability in posterior predicted survival from the CP frailty model than the latent class CP frailty mixture model. Both frailty models tend to overestimate hazard when applied to the Partners PrEP data (Figures 3.9 and 3.10). Survival is underestimated in all trial arms, including the oral placebo arm, implying that the probability of exposure or level of risk is overestimated rather than the effectiveness of the candidate interventions being overestimated. In analysis of the VOICE data (Figures 3.11 and 3.12), both frailty models overestimated hazard at the beginning of the trial and underestimated hazard at the end. This trend is more pronounced in the oral intervention arms, suggesting some time-dependence in effectiveness. Such a trend would be expected for beneficial interventions if adherence was better at the start of the trial than at the end.

Plots showing the prior and posterior distributions for all parameters in both models used to analysis Partners PrEP and VOICE trial data are given in Figures 3.13 - 3.37. The posterior distribution of most parameters differ substantially from their priors. Notably, the posterior distributions for class-specific values of η from the mixture models, shown in the right-hand panel of Figures 3.14 and 3.28, vary considerably from their priors while the site-specific posteriors of η from the single CP frailty models, 3.20 and 3.35, are more similar to their priors. Finally, following our discussion in Section 3.2.1 about the interpretation of cumulative link model coefficients when using a logistic link function, we see that coefficients for characteristics associated with lower risk (for example, partner's baseline CD4 count in Partners PrEP or being married/cohabitating in VOICE) have posterior distributions centered below zero while coefficients for characteristics associated with higher risk (for example, no children in Partners PrEP or age at or below 25 years in VOICE) have posterior distributions centered above zero.

3.5 Discussion

We have proposed a novel compound Poisson mixture frailty model for univariate time-to-event data that exploits information available in covariates in order to account for heterogeneity in risk. Our approach uses ordinal latent class regression to identify participants with similar risk characteristics. Individuals within a class then share a frailty distribution. The

compound Poisson distribution was used to model frailty within latent classes as it provides a mechanistically-motivated model for heterogeneity in risk and allows for some participants to have no risk of an event. Under a compound Poisson frailty framework, an individual's overall risk is the result of independent, competing exposure processes, the number and nature of which remain unobserved. In the proposed model, the number of exposure processes for individuals and the magnitude associated with distinct exposure processes follow separate distributions for each latent class. Frailty distributions are stochastically ordered so that average frailty is smaller in lower risk classes and greater in higher risk classes, yet the frailties of individuals within each class vary so that an individual in a lower risk class may have greater risk than an individual in a higher risk class and vice-versa. In this framework, even individuals in high risk classes can have no risk of an event. The resulting model enables estimation of individual-level hazard ratios for interventions or conditions of interest, instead of population-averaged hazard ratio estimates, like those obtained from the Cox proportional-hazards model.

With the proposed model, we estimated the individual-level effectiveness of ARV-based interventions in two recently completed HIV prevention trials: Partners PrEP and VOICE. For some interventions, the estimate of effectiveness for an individual at risk obtained from the proposed model was greater than the population-averaged estimate obtained by the Cox model, as we would expect when there was heterogeneity in risk of an event. Effectiveness estimates of particular interventions are not similar across trials; the estimated effectiveness of both oral TDF and TDF-FTC are higher in the analysis of Partners PrEP data than in that of the VOICE trial data. These discrepancies were anticipated given that treatment adherence was very high in Partners PrEP and quite poor among participants in VOICE.

In addition to providing an individual-level estimate of hazard ratios, the proposed approach is particularly appealing when there is interest in identifying groups of participants with similar levels of risk. For example, identifying individuals with very high risk of an event could be helpful in targeting interventions and planning future studies. Additionally, the proposed model could be easily extended to estimate hazard ratios separately for each risk class. In our application, we demonstrated that the proposed model clearly distinguished between low, medium, and high risk classes. While we used covariates that were significant

predictors of an event in a univariate model, Bayesian model selection methods could be used to choose risk-related covariates for the latent classification model from a pool of candidate covariates [14, 22]. The number of classes could also be determined with Bayesian model selection methods.

In our application we also found that the proposed model may not be ideal in a clinical trial setting where adherence is expected to be associated with risk of an event. If better adherence is associated with lower risk, as may have been the case in the VOICE trial, the model may underestimate the probability of any risk for adherent participants. The resulting estimate of intervention effectiveness would then underestimate the true effectiveness. Further research is necessary to better understand this relationship. In particular, it would be interesting to examine model performance when adherence is partially observed, as was the case in the VOICE trial, so that effectiveness among those who adhere can be modeled through various weighting methods (for example, marginal structural models [24, 44]).

In the proposed model, participants within a latent class share a value for both frailty parameters, ρ and ν. It is possible, however, to model latent risk classes for these parameters separately such that individuals are classified twice. This modification would be appropriate if information were available on covariates that predicted the level of risk given exposure, η, but were not associated with the probability of exposure or number of exposure processes, ρ. The decision to model separate latent classes depends on the application and data available. A shared latent class approach was presented here for simplicity in presentation.

While the proposed model does offer a novel approach to incorporating covariate data in the analysis of time-to-event data, it does not address another limitation of univariate frailty models: individuals' frailties are assumed to be constant over time. This constraint is necessary because, in the absence of multiple event times, information is not available to model time-varying frailty. In our application, frailty parameters are determined based on baseline covariates. But, we expect the values of some of these covariates to change over the study period, and the proposed model has no way or reflecting that change. Certainly, allowing individuals' frailties to vary over time, in concert with changes in risk-related covariates, would improve heterogeneity modeling. It is possible that time-varying covariates could provide information about changes in an individual's frailty over time, and that, even

with a single event time, the proposed model could be extended to use these covariates to model time-varying frailties. On the other hand, given that the latent class model for constant frailty presented here did not provide more efficient estimates than the single compound Poisson frailty model, extending the model to allow for time-varying frailties would likely yield a negligible improvement in estimation. As an alternative to time-varying frailty models, we may also consider how to incorporate the proposed compound Poisson mixture frailty model into a joint model for longitudinal and survival data [12, 53].

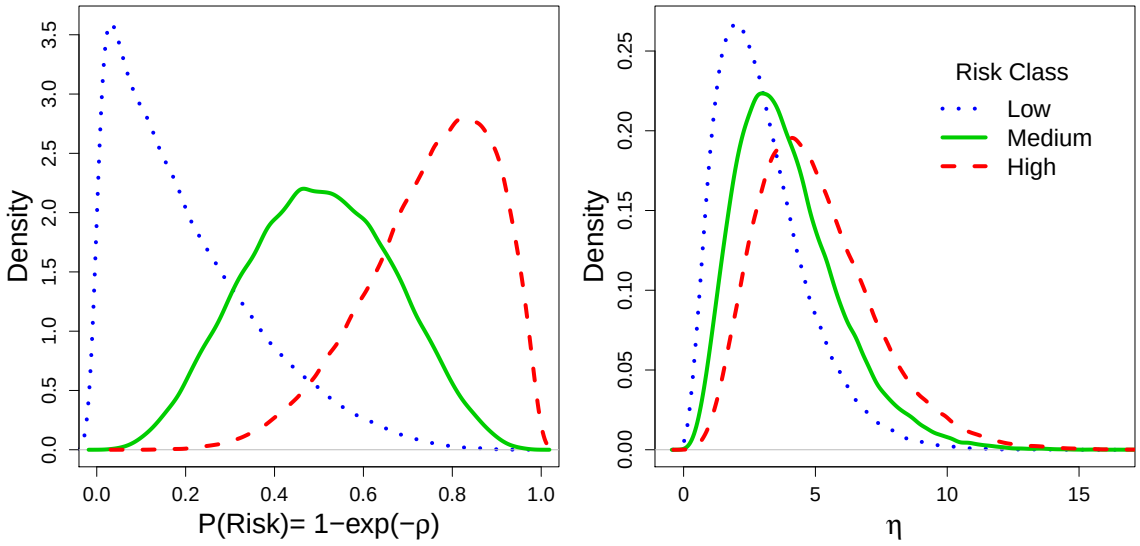

Figure 3.1: Prior distributions for the proportion at risk, $1 - \exp(-\rho)$, (left panel) and the shape parameter associated with each exposure process, η, (right panel), for three latent risk classes.

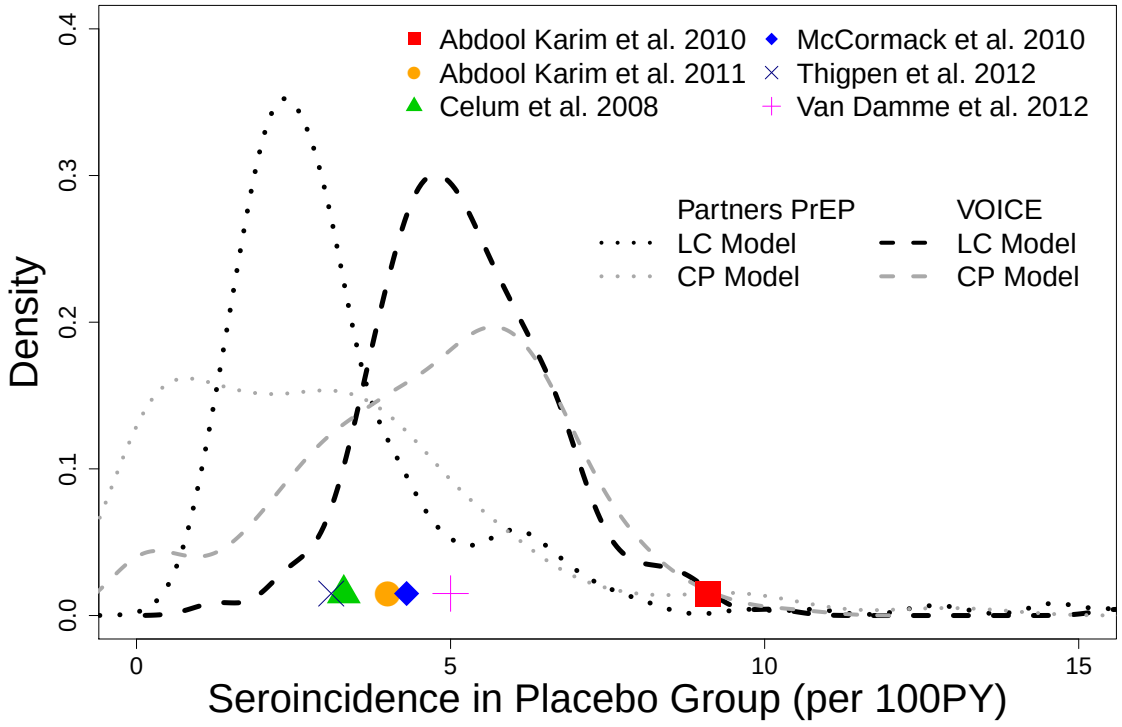

Figure 3.2: Prior distribution of seroincidence in the placebo arm for Partners PrEP (dotted) and VOICE (dashed) analysis compared to observed incidence in previous HIV prevention trials.

	n	Incidence	Classification Probabilities		
Covariate	(%)	(per 100PY)	Low	Medium	High
All	4,945 (100%)	5.1	39%	28%	28%
Study Site Location					
Durban, South Africa	3,329 (67%)	7.3	7.3%	41%	51%
Johannesburg, South Africa	670 (14%)	2.1	15%	64%	19%
Kampala, Uganda	320 (6.5%)	1.7	18%	72%	8.1%
Harare, Zimbabwe	626 (13%)	0.45	30%	67%	2.7%
Age (years)					
≤ 25	2,857 (58%)	7.7	6.2%	43%	50%
>25	2,088 (42%)	3.5	22%	56%	21%
# Previous live births					
None	746 (15%)	6.9	5.2%	42%	51%
1	2,138 (43%)	6.3	8.3%	45%	45%
≥ 2	2,061 (42%)	3.4	20%	54%	26%
Married and/or Cohabitating					
Yes	1,631 (33%)	1.5	29%	60%	10%
No	3,314 (67%)	7.2	5.1%	43%	51%
Gonorrhea test at baseline					
Positive	155 (3.1%)	13.4	5.2%	35%	58%
Negative/ None done	4,790 (97%)	4.9	13%	49%	37%
Primary partner has other sexual partners					
Yes	757 (15%)	4.6	15%	52%	32%
No	1,218 (25%)	3.3	17%	52%	28%
Don't know/ No primary partner	2,990 (60%)	6.0	10%	46%	44%

Table 3.8: Median posterior classification probabilities across covariate values for **VOICE** trial.

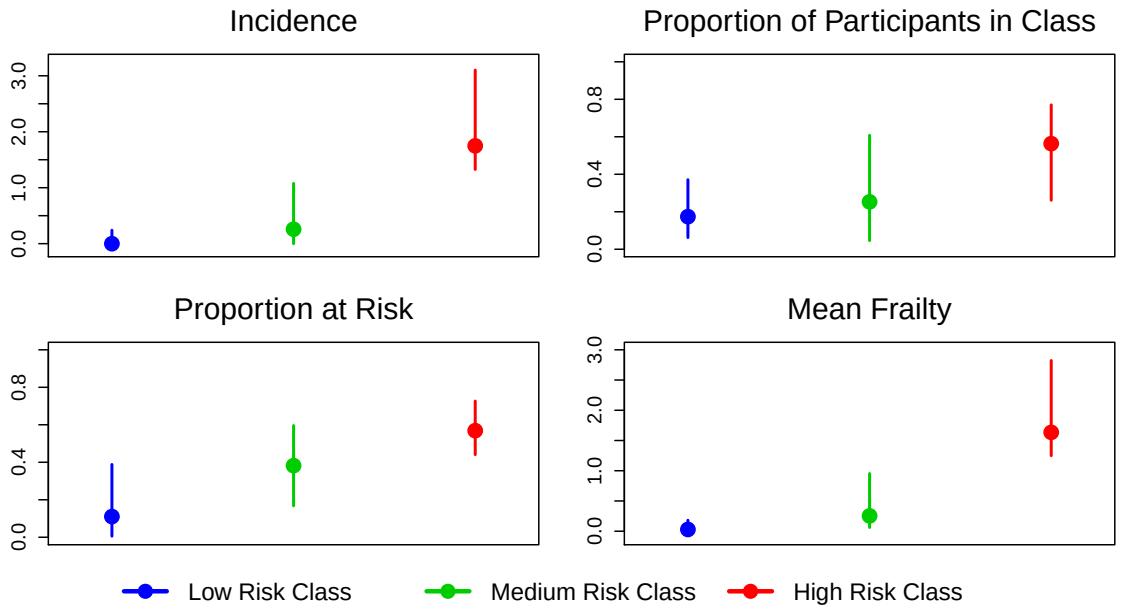

Figure 3.3: Posterior medians (points) and 95% credible intervals (lines) of summary measures of latent risk classes for **Partners PrEP** analysis.

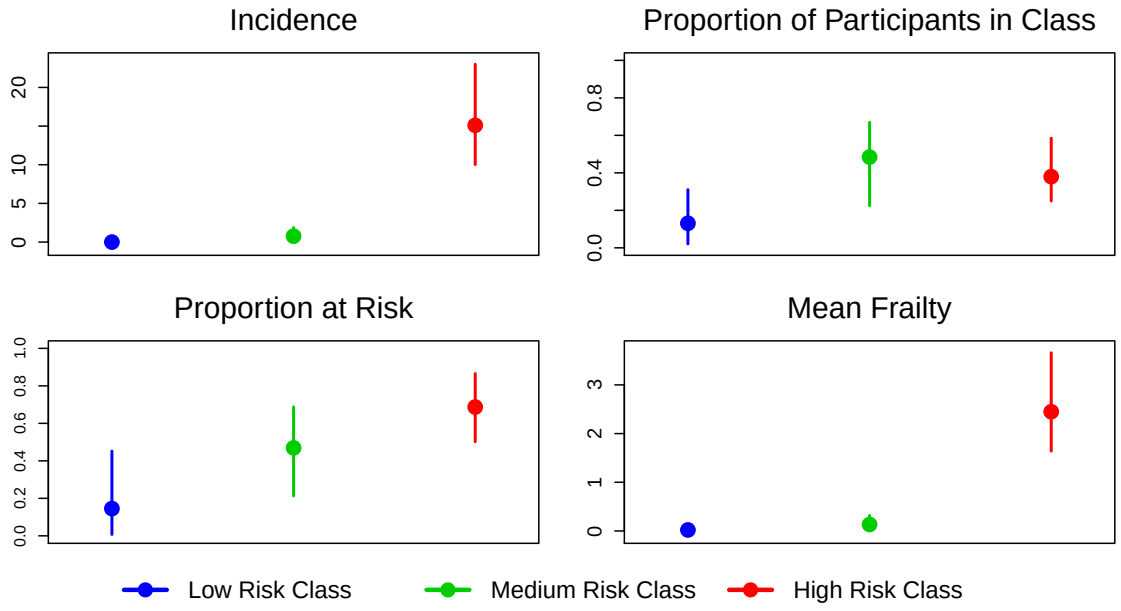

Figure 3.4: Posterior medians (points) and 95% credible intervals (lines) of summary measures of latent risk classes for **VOICE** analysis.

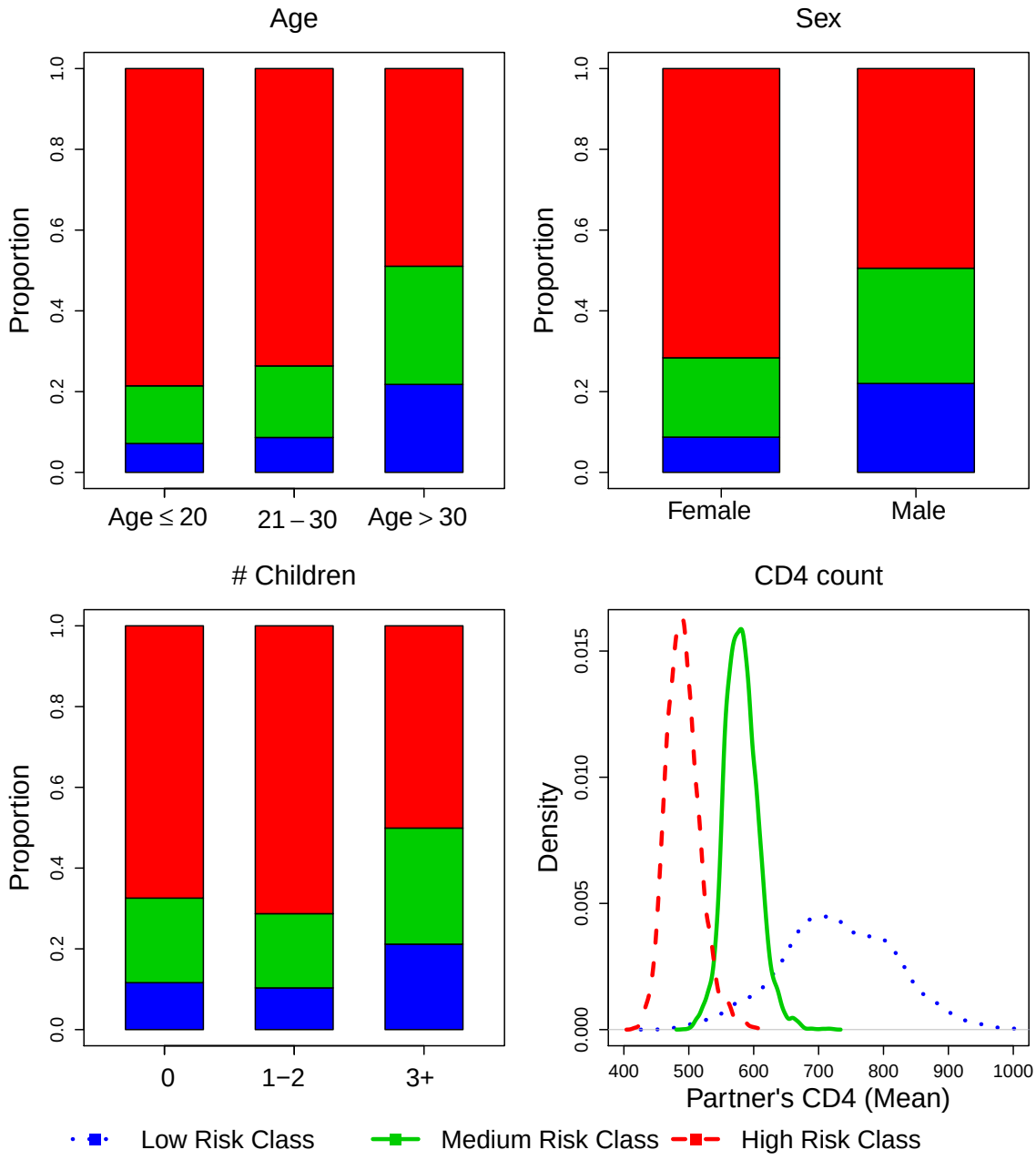

Figure 3.5: Median posterior classification probabilities across covariate values for **Partners PrEP** trial.

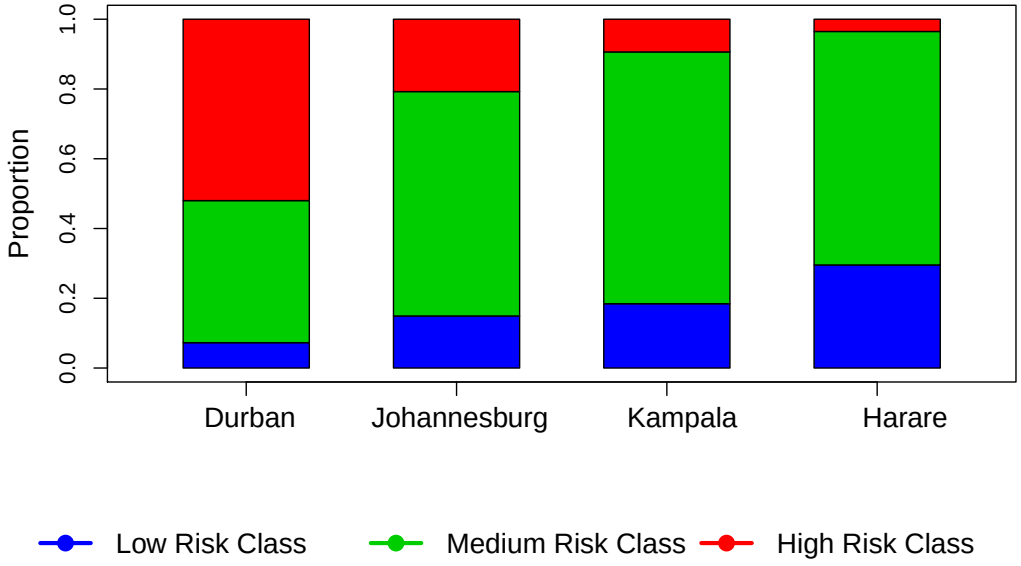

Figure 3.6: Median posterior classification probabilities across covariate values for **VOICE** trial.

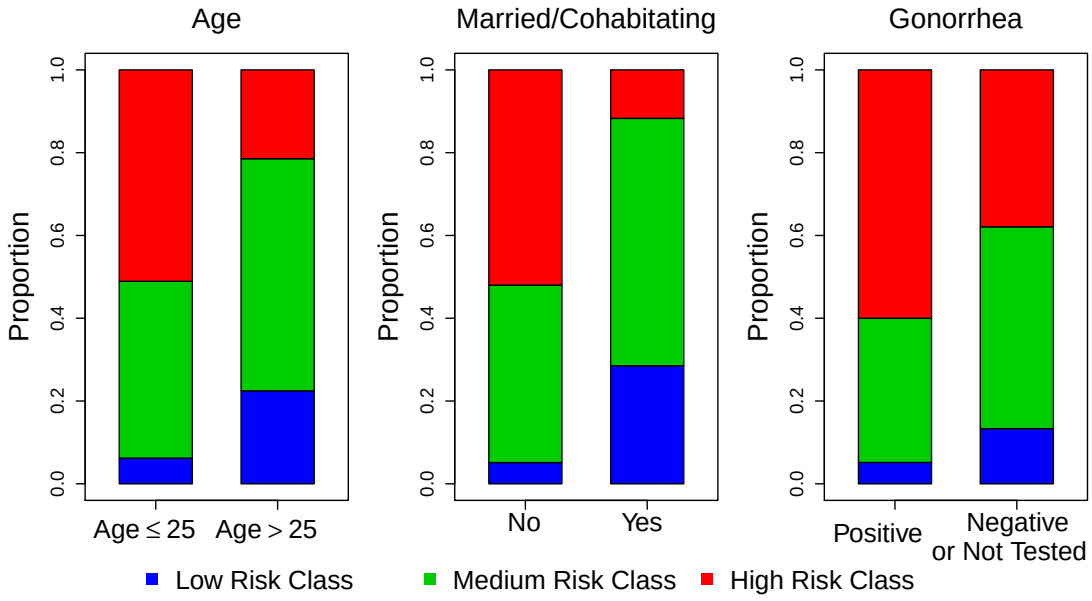

Figure 3.7: Median posterior classification probabilities across covariate values for **VOICE** trial.

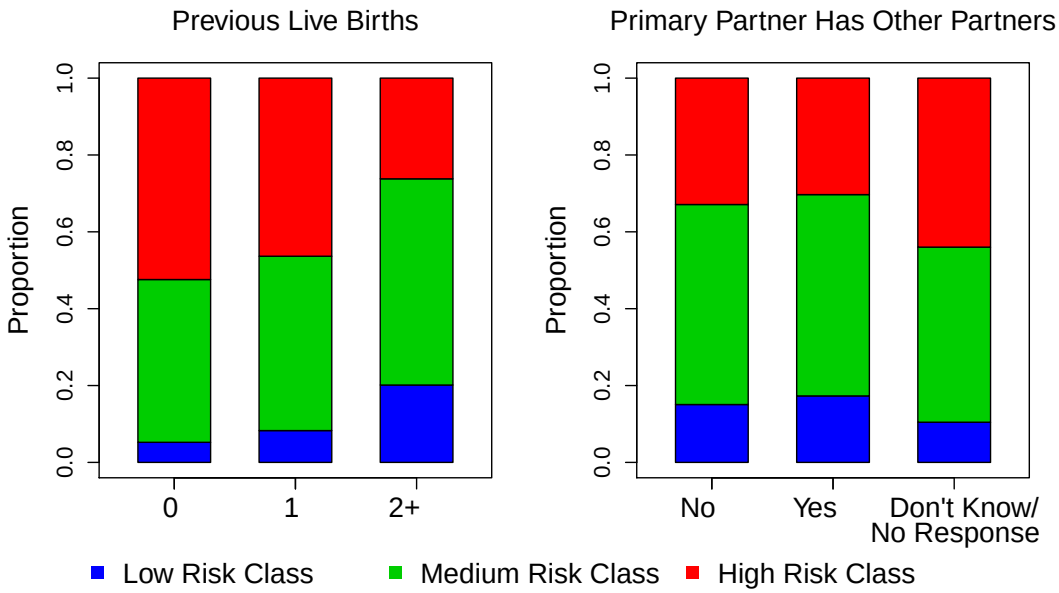

Figure 3.8: Median posterior classification probabilities across covariate values for **VOICE** trial.

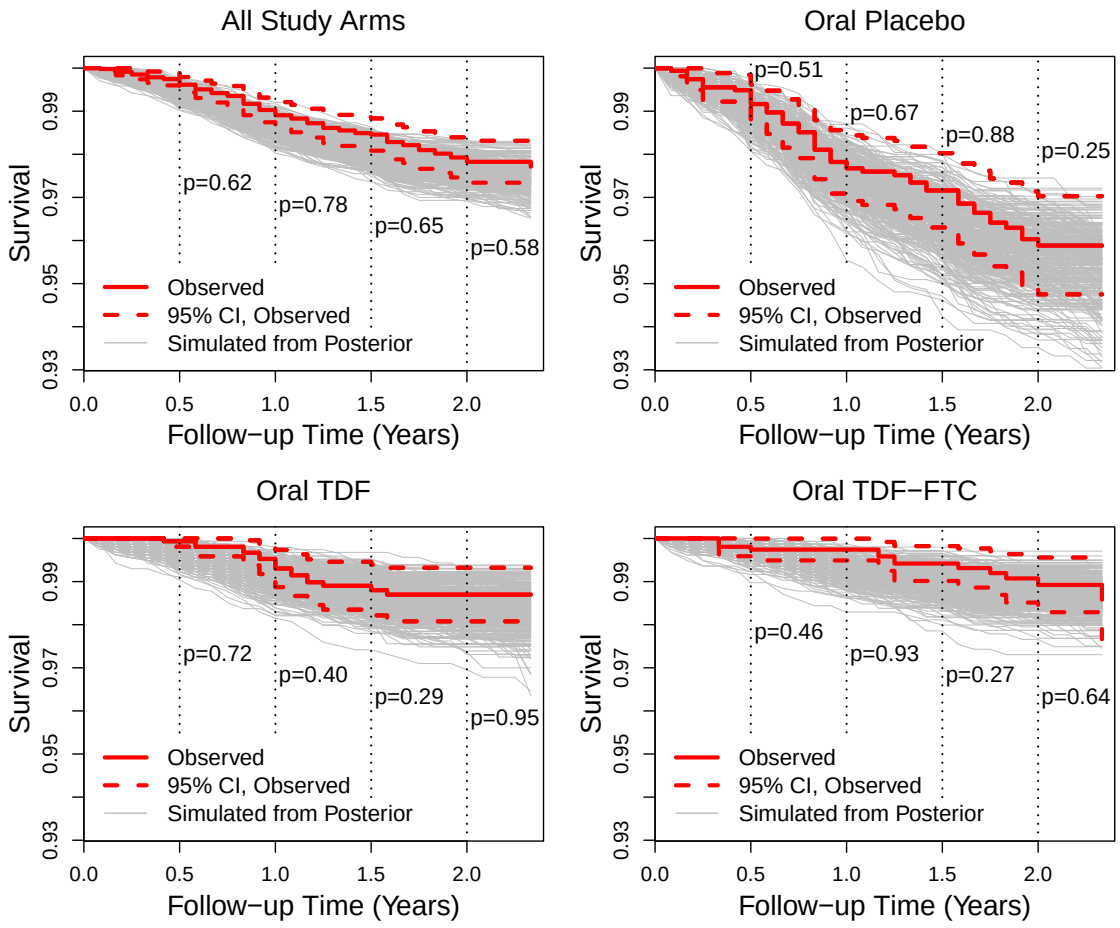

Figure 3.9: Kaplan Meier curves for observed survival in the **Partners PrEP** trial (bold, solid lines) and surrounding 95% confidence intervals (dashed lines) with survival curves generated under the posterior predictive distribution of the **latent class compound Poisson mixture frailty** model. Bayesian p-values comparing observed to predicted survival are given at six-month intervals.

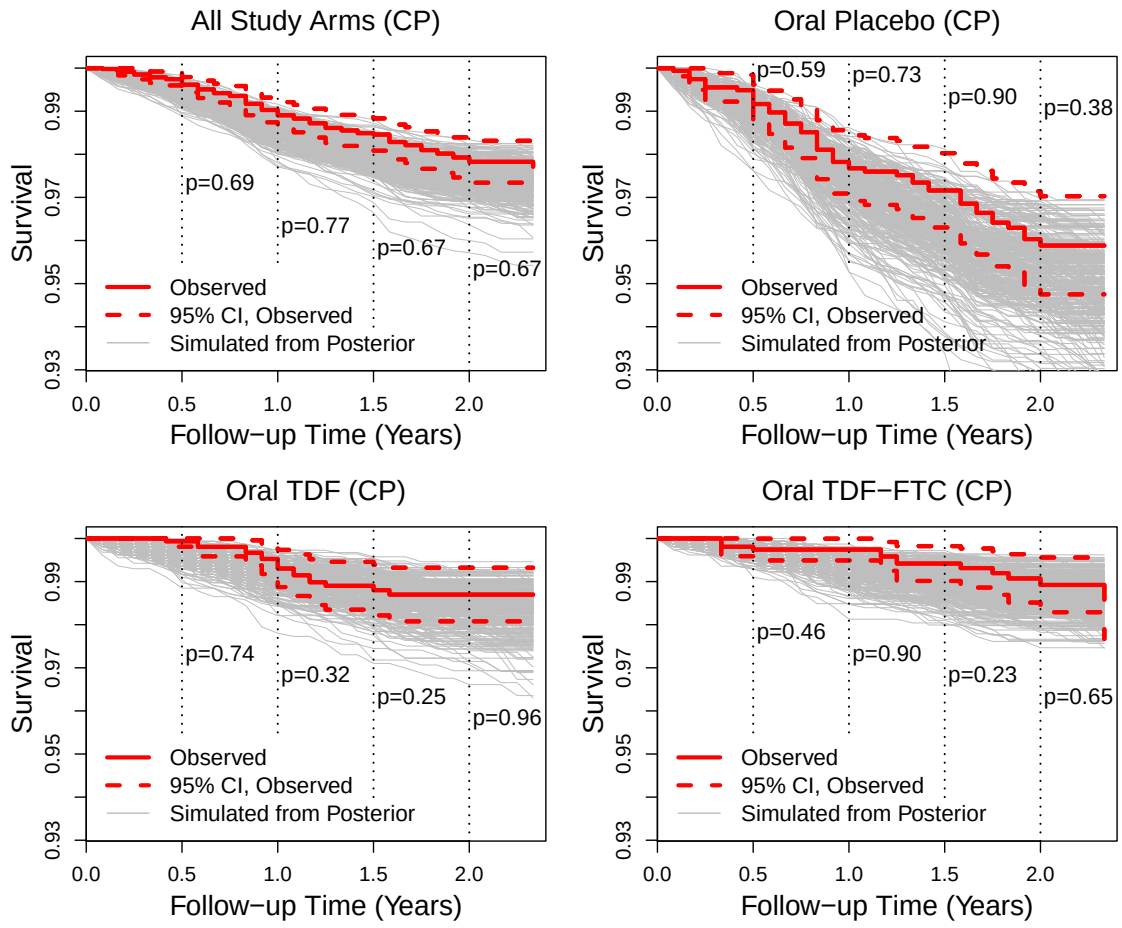

Figure 3.10: Kaplan Meier curves for observed survival in the **Partners PrEP** trial (bold, solid lines) and surrounding 95% confidence intervals (dashed lines) with survival curves generated under the posterior predictive distribution of the **country-stratified compound Poisson frailty** model. Bayesian p-values comparing observed to predicted survival are given at six-month intervals.

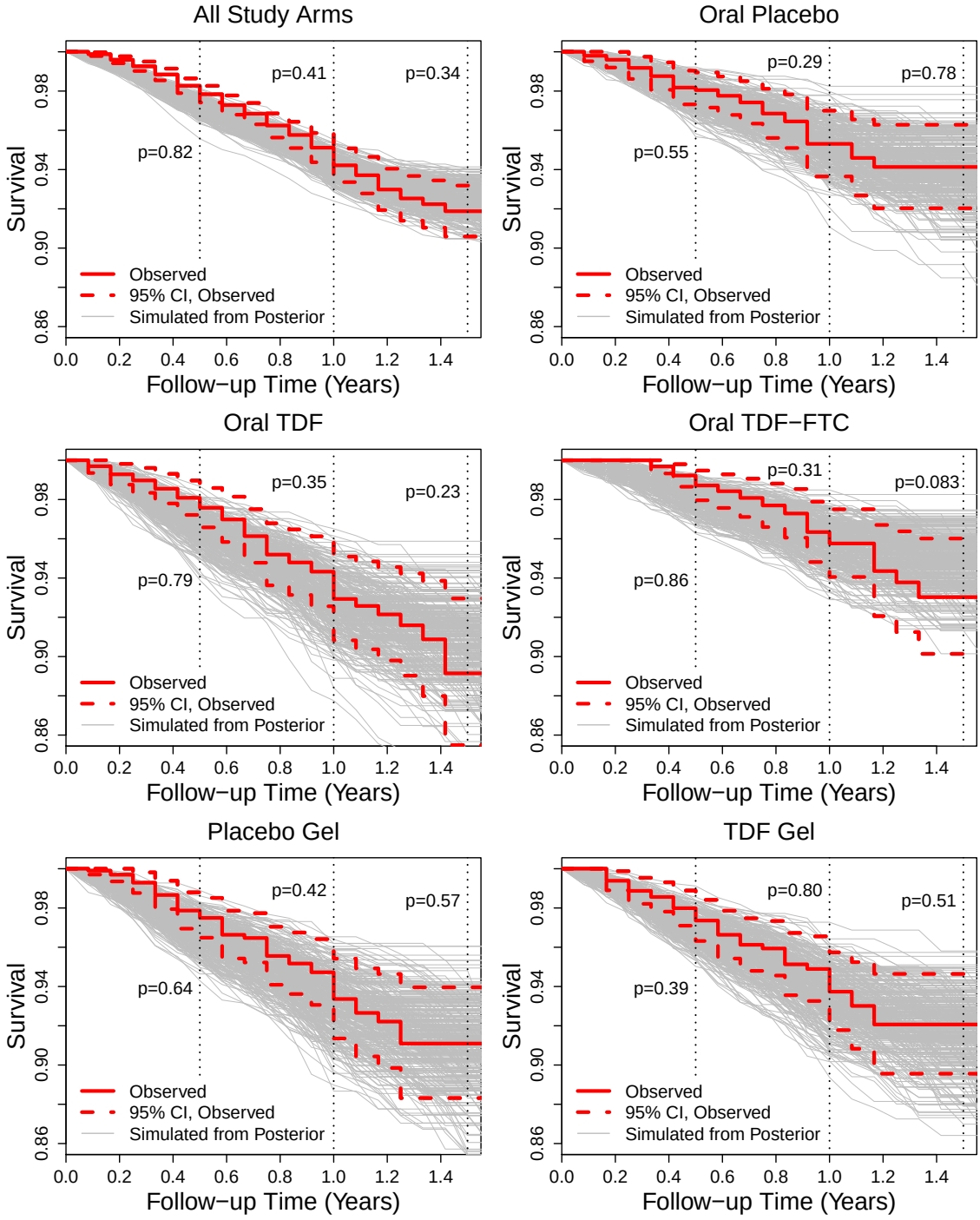

Figure 3.11: Kaplan Meier curves for observed survival in the **VOICE** trial (bold, solid lines) and surrounding 95% confidence intervals (dashed lines) with survival curves generated under the posterior predictive distribution of the **latent class compound Poisson mixture frailty** model. Bayesian p-values comparing observed to predicted survival are given at six-month intervals.

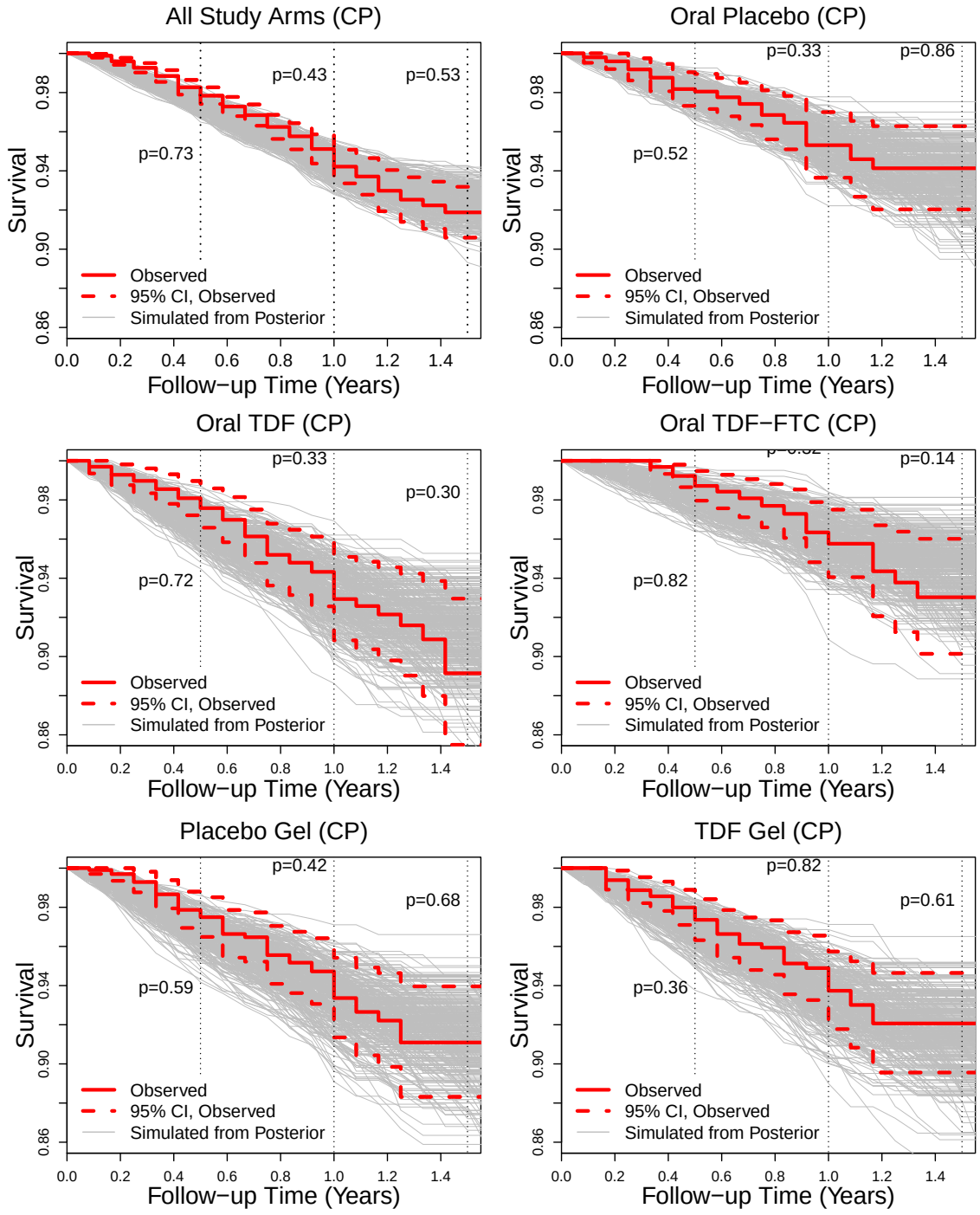

Figure 3.12: Kaplan Meier curves for observed survival in the **VOICE** trial (bold, solid lines) and surrounding 95% confidence intervals (dashed lines) with survival curves generated under the posterior predictive distribution of the **site location-stratified compound Poisson frailty** model. Bayesian p-values comparing observed to predicted survival are given at six-month intervals.

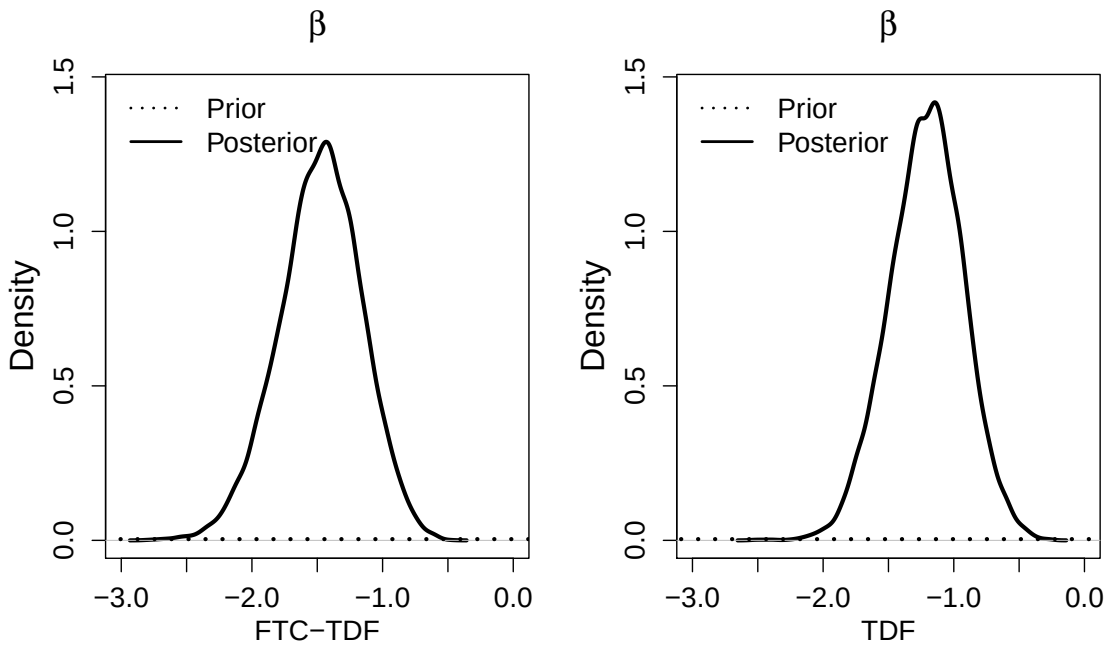

Figure 3.13: Prior and posterior distributions of log hazard ratio vector, $\boldsymbol{\beta}$, from the **latent class compound Poisson mixture frailty** model analysis of **Partners PrEP** data.

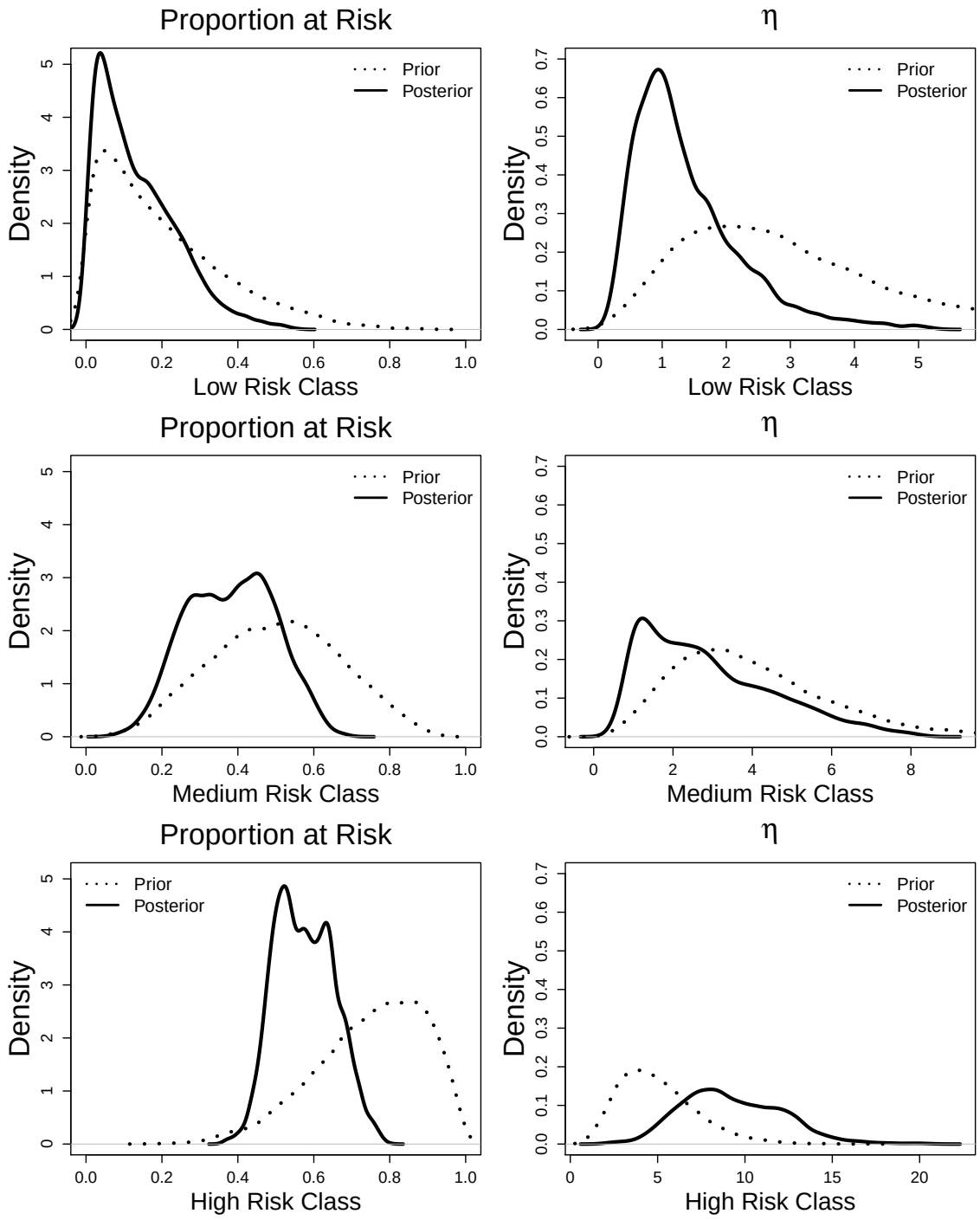

Figure 3.14: Prior and posterior distributions of the proportion at risk and η for each latent class from the **latent class compound Poisson mixture frailty** model analysis of **Partners PrEP** data.

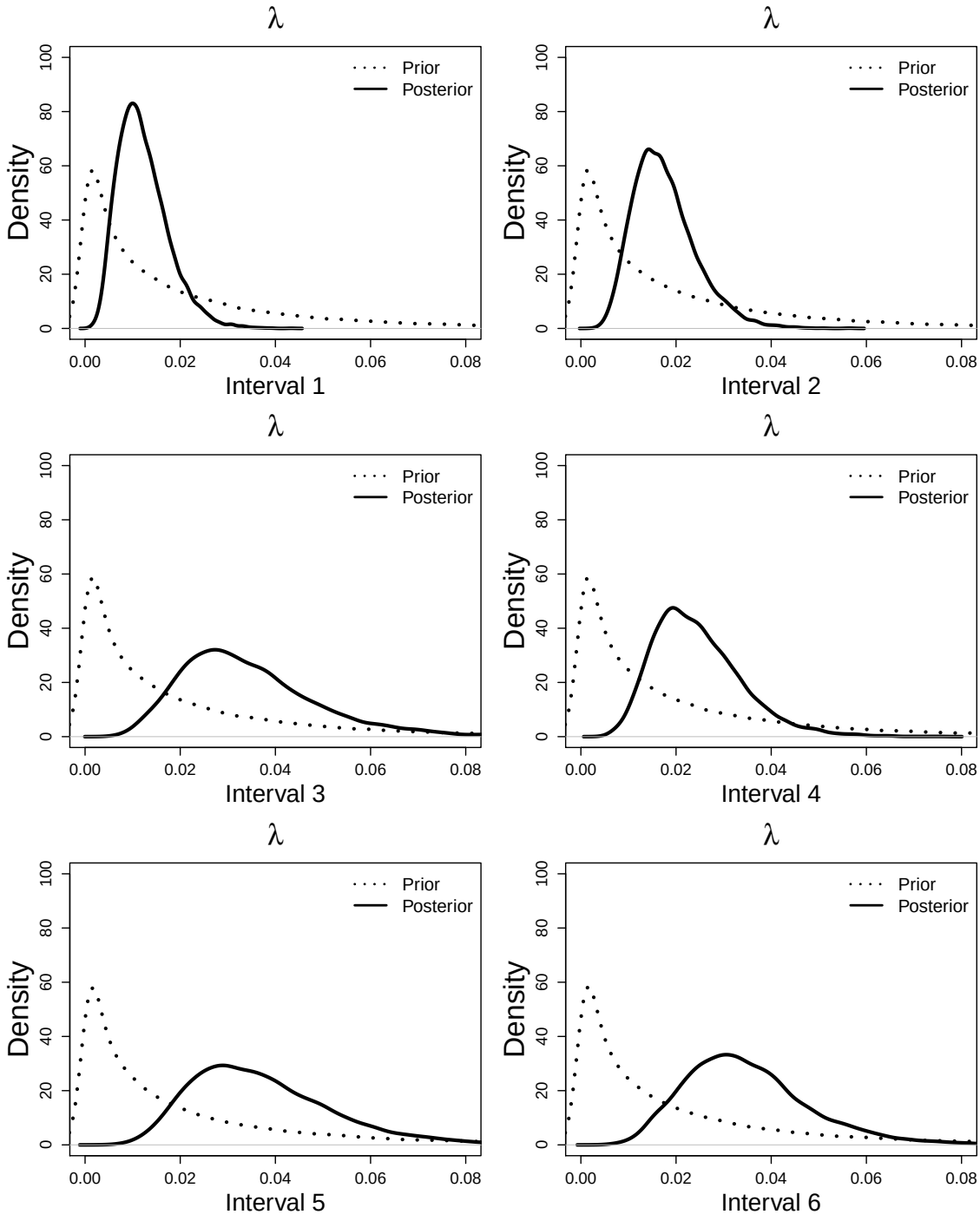

Figure 3.15: Prior and posterior distributions of piecewise constant baseline hazard from the **latent class compound Poisson mixture frailty** model analysis of **Partners PrEP** data.

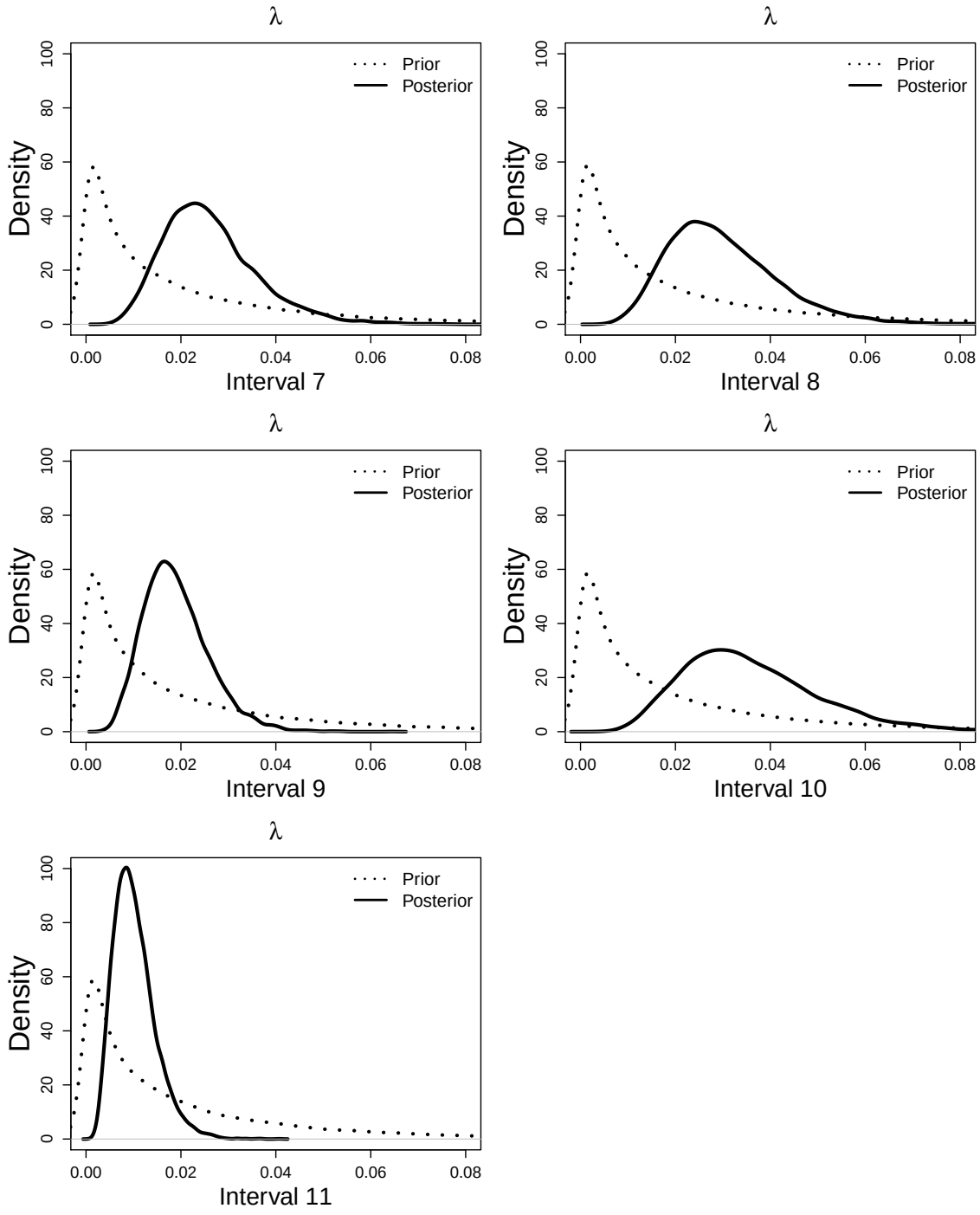

Figure 3.16: Prior and posterior distributions of piecewise constant baseline hazard from the **latent class compound Poisson mixture frailty** model analysis of **Partners PrEP** data.

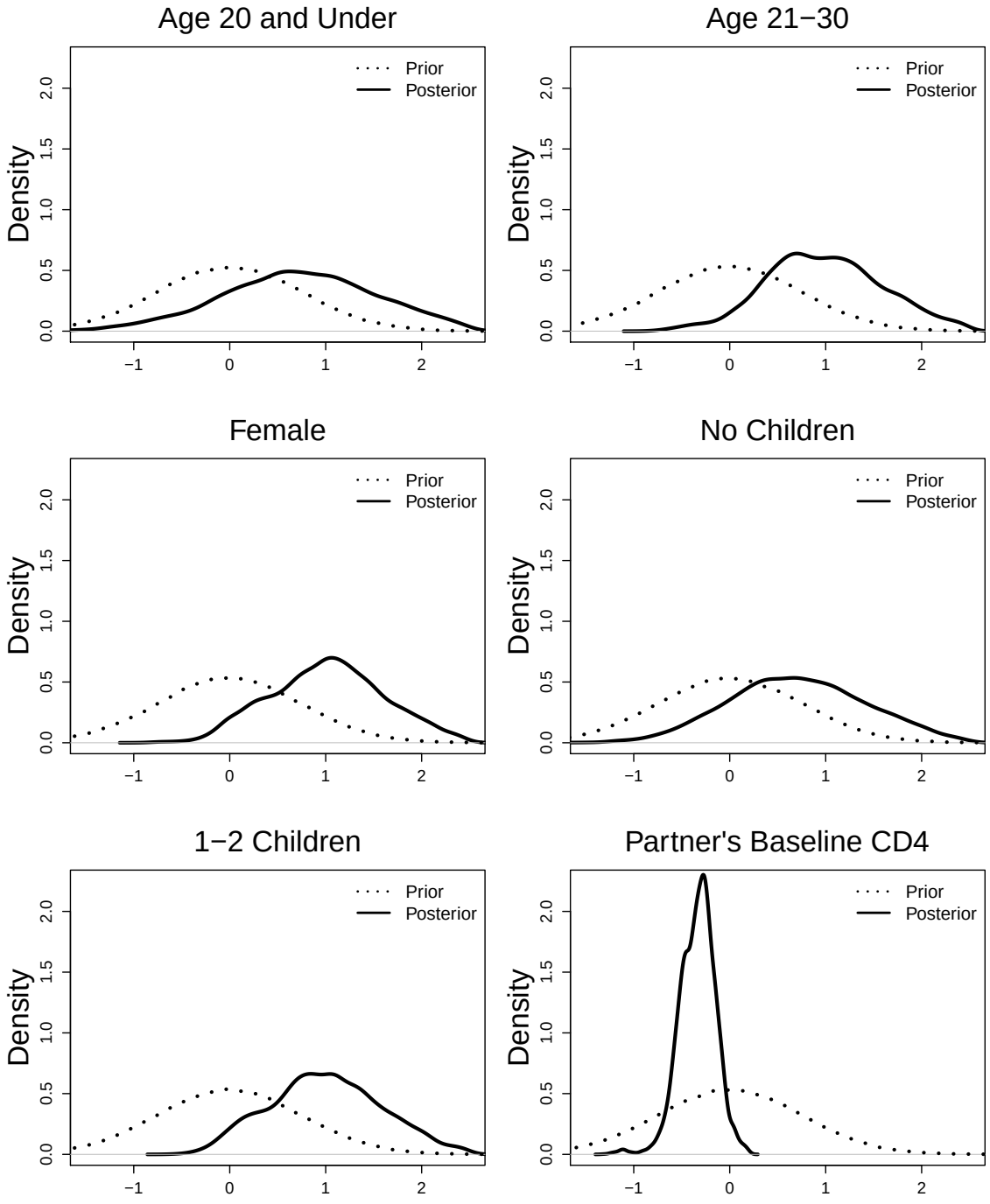

Figure 3.17: Prior and posterior distributions for proportional odds model covariate effects $\boldsymbol{\gamma}^*$ from the **latent class compound Poisson mixture frailty** model analysis of **Partners PrEP** data.

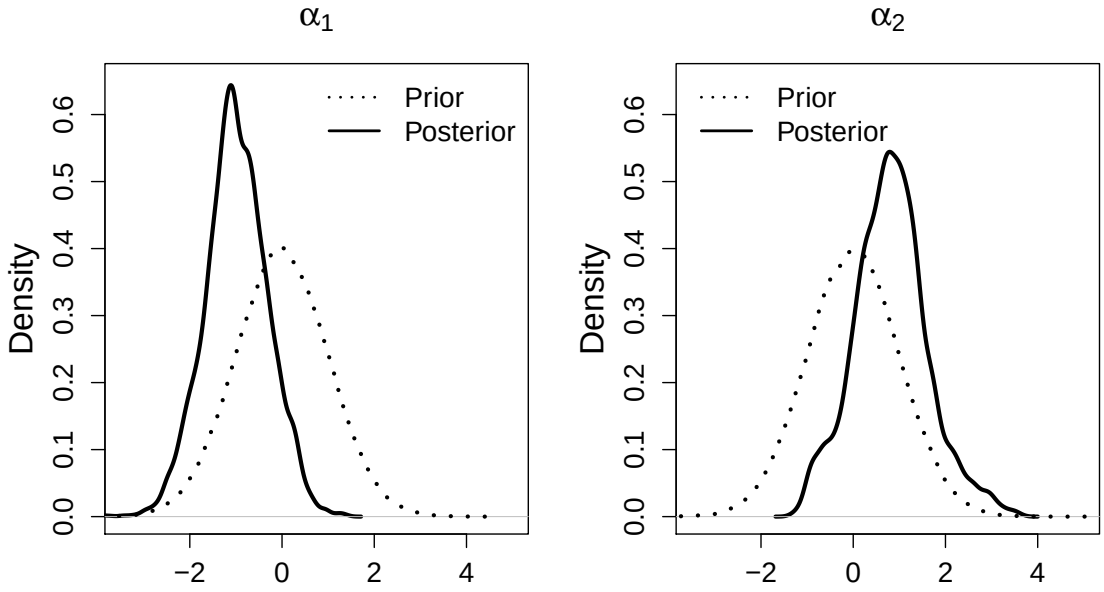

Figure 3.18: Prior and posterior distributions for proportional odds model intercepts α_1 and α_2 from the **latent class compound Poisson mixture frailty** model analysis of **Partners PrEP** data.

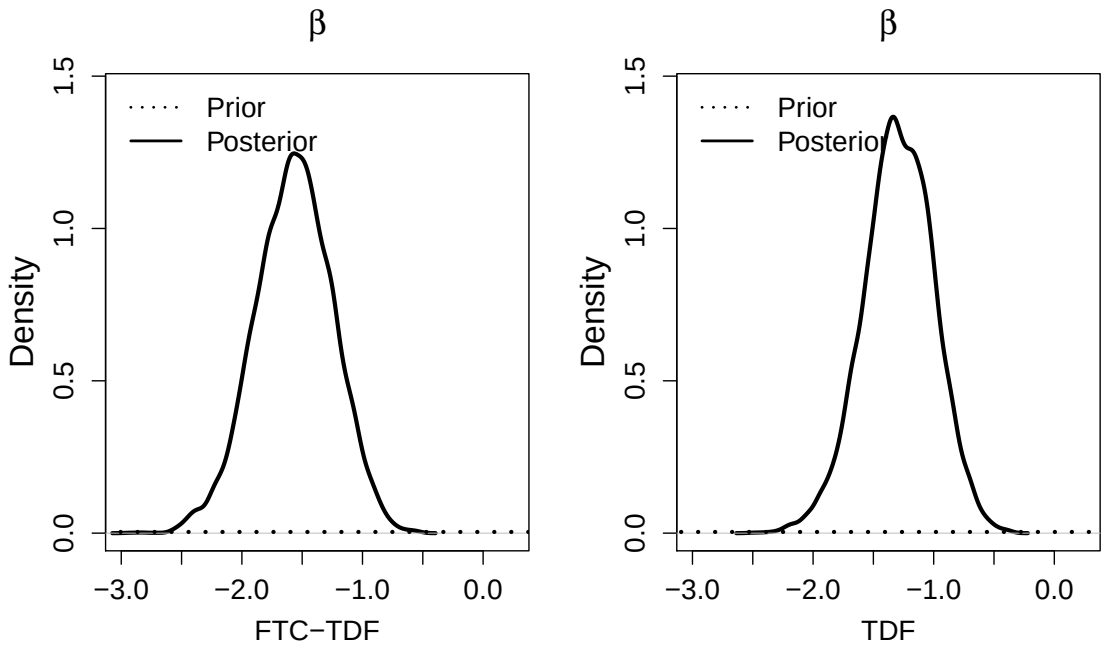

Figure 3.19: Prior and posterior distributions for log- hazard ratios from the **country-stratified compound Poisson frailty** model analysis of **Partners PrEP** data.

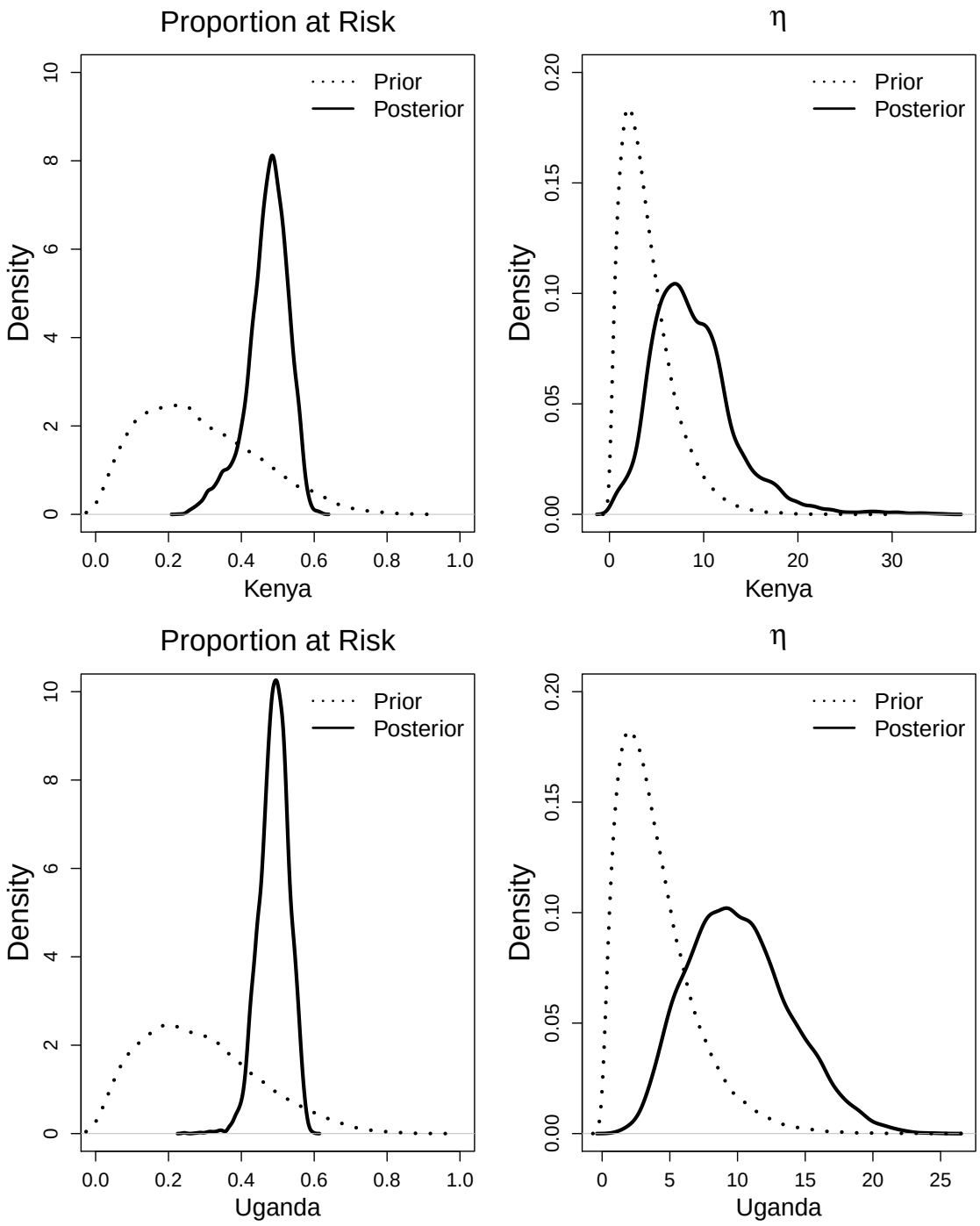

Figure 3.20: Prior and posterior distributions for proportion at risk and η from the **country-stratified compound Poisson frailty** model analysis of **Partners PrEP** data.

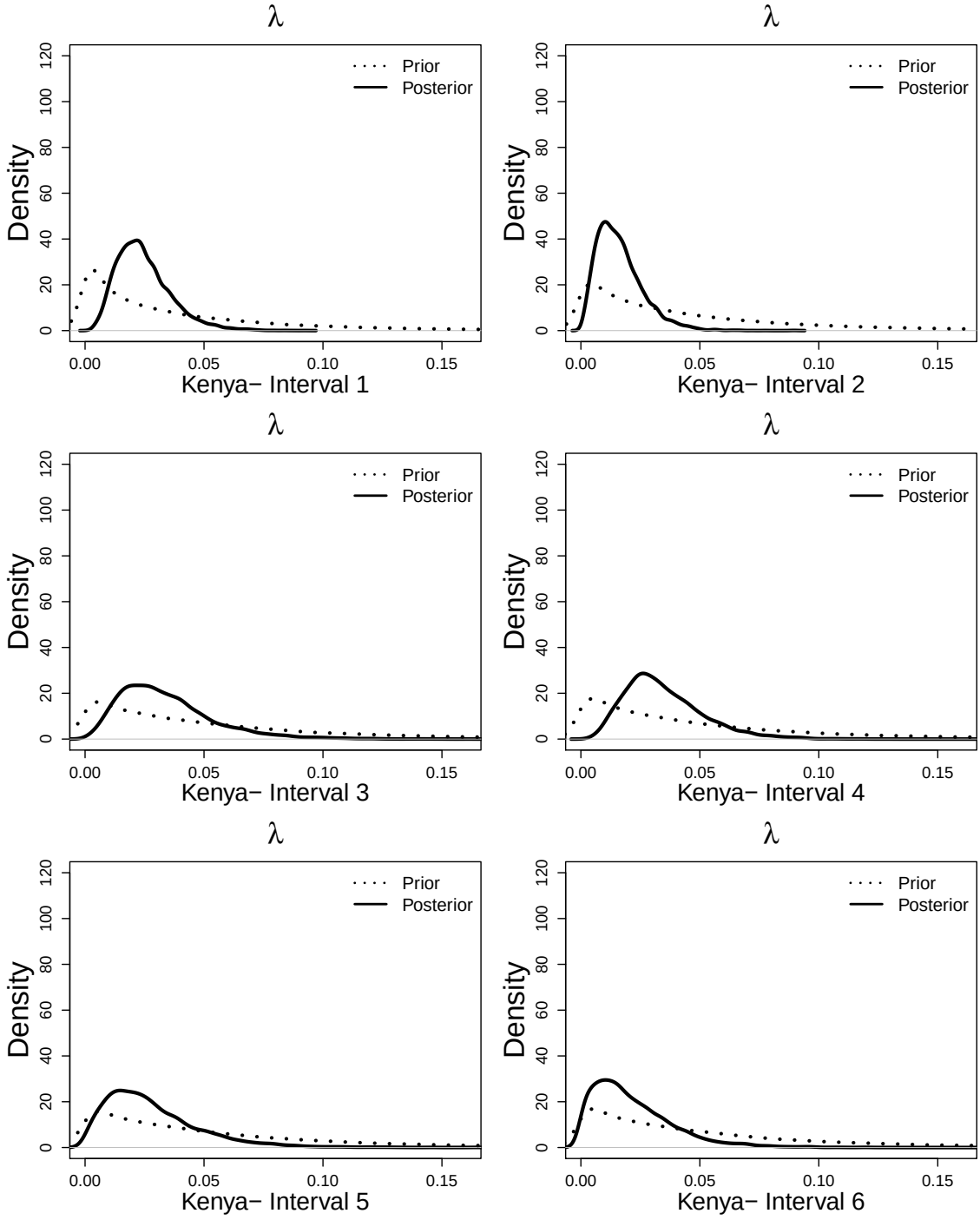

Figure 3.21: Prior and posterior distributions for piecewise baseline hazards for Kenyan sites from the **country-stratified compound Poisson frailty** model analysis of **Partners PrEP** data.

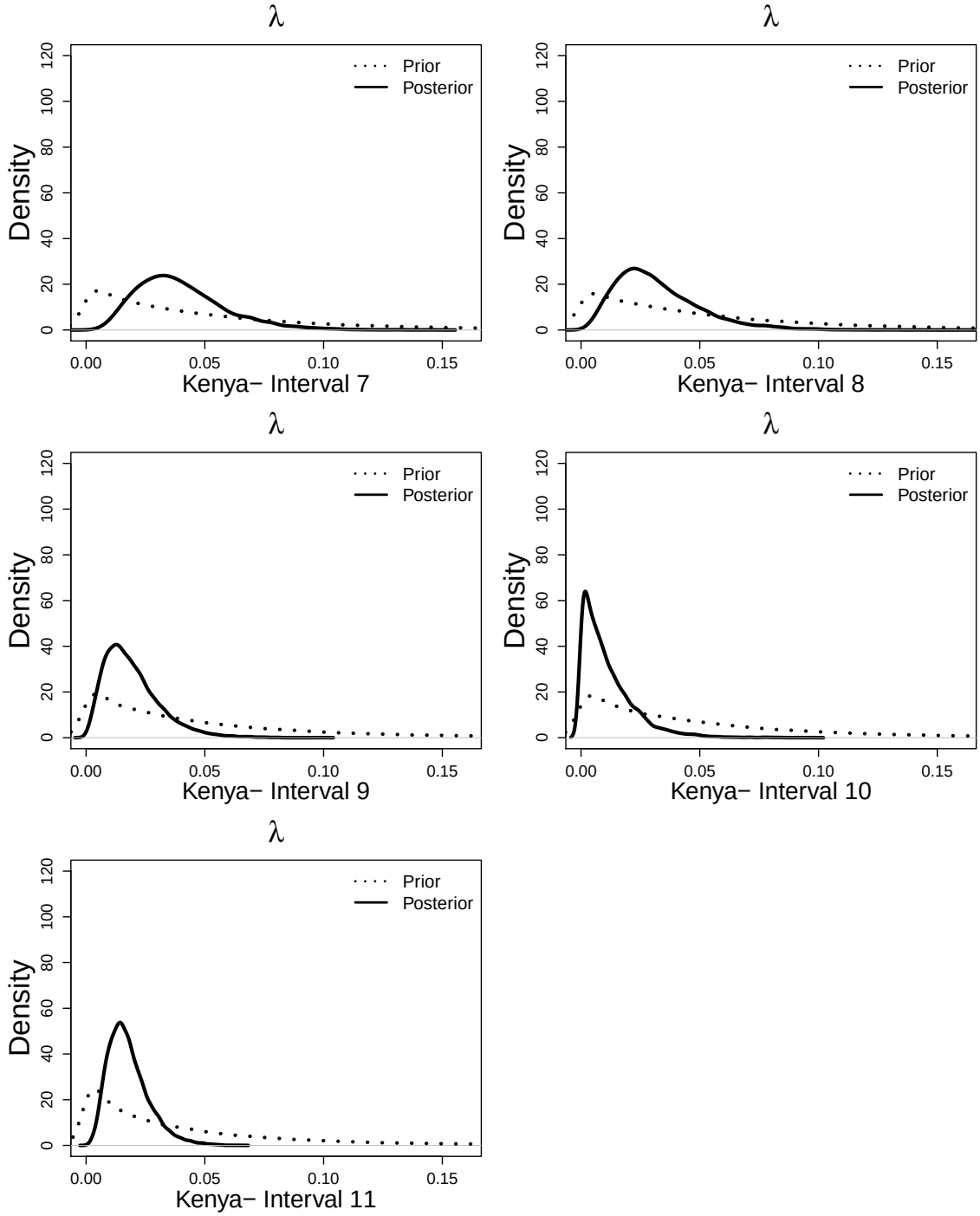

Figure 3.22: Prior and posterior distributions for piecewise baseline hazards for Kenyan sites from the **country-stratified compound Poisson frailty** model analysis of **Partners PrEP** data.

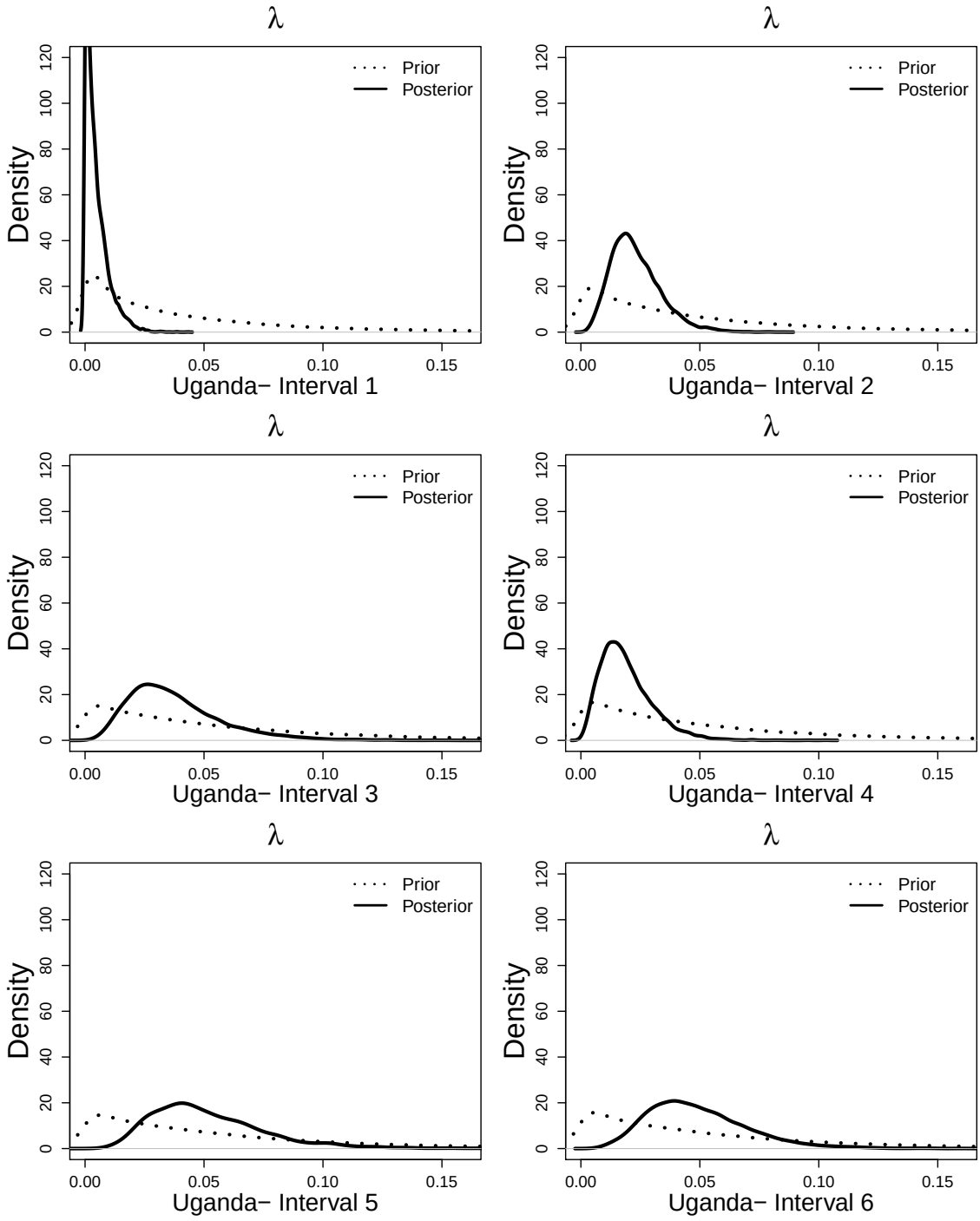

Figure 3.23: Prior and posterior distributions for piecewise baseline hazards for Ugandan sites from the **country-stratified compound Poisson frailty** model analysis of **Partners PrEP** data.

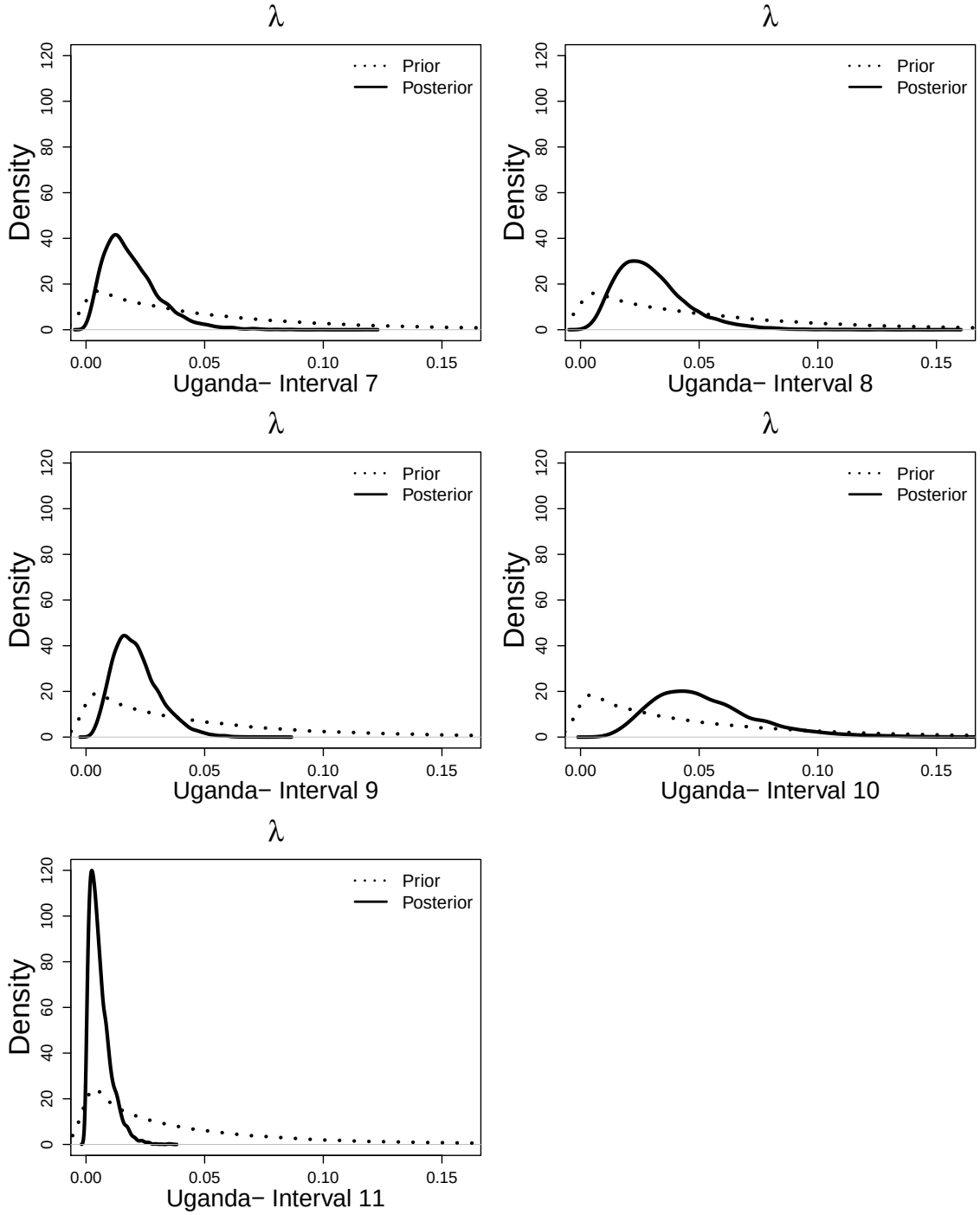

Figure 3.24: Prior and posterior distributions for piecewise baseline hazards for Ugandan sites from the **country-stratified compound Poisson frailty** model analysis of **Partners PrEP** data.

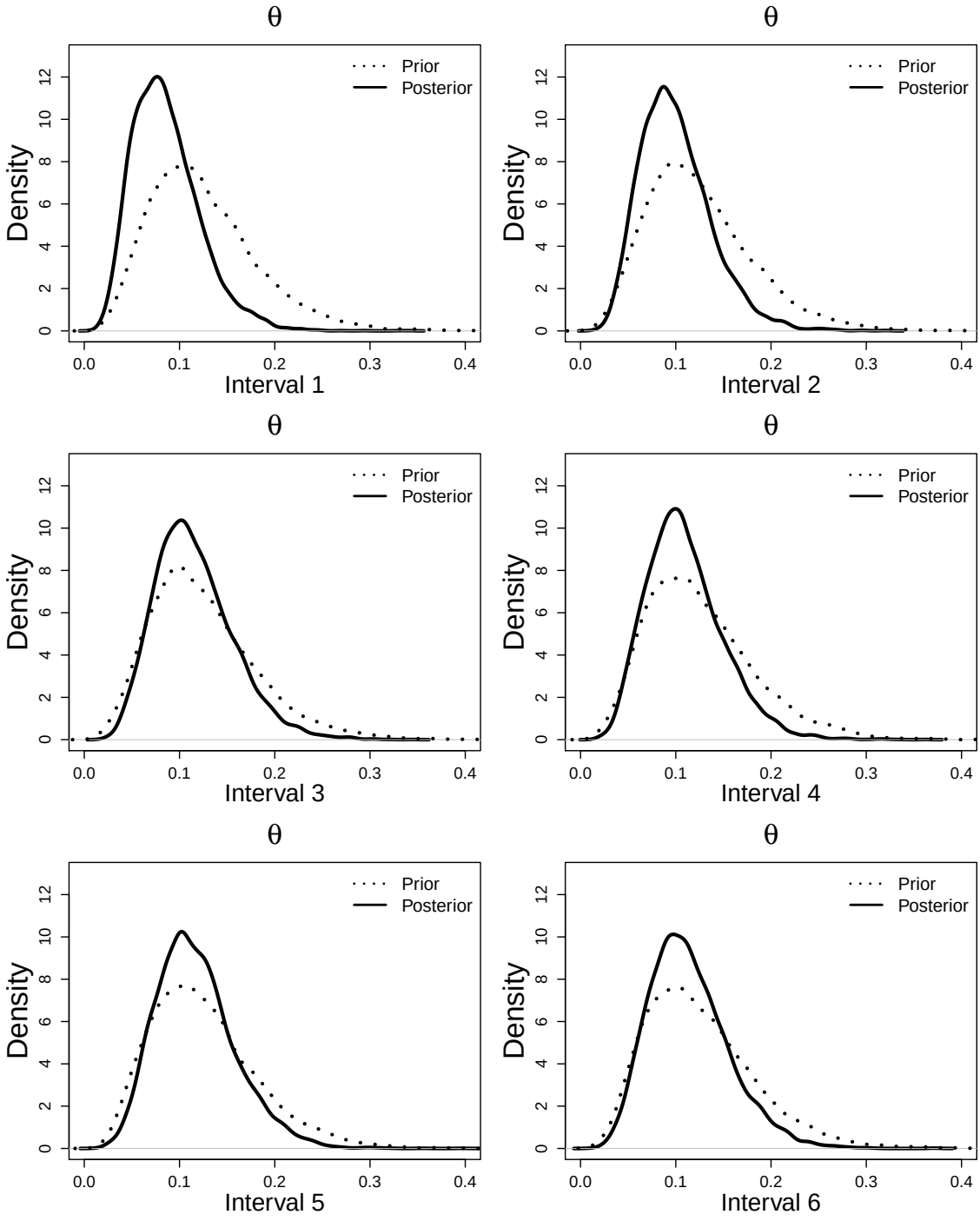

Figure 3.25: Prior and posterior distributions for the piecewise baseline hazard function for at-risk participants, θ, from the **country-stratified compound Poisson frailty** model analysis of **Partners PrEP** data.

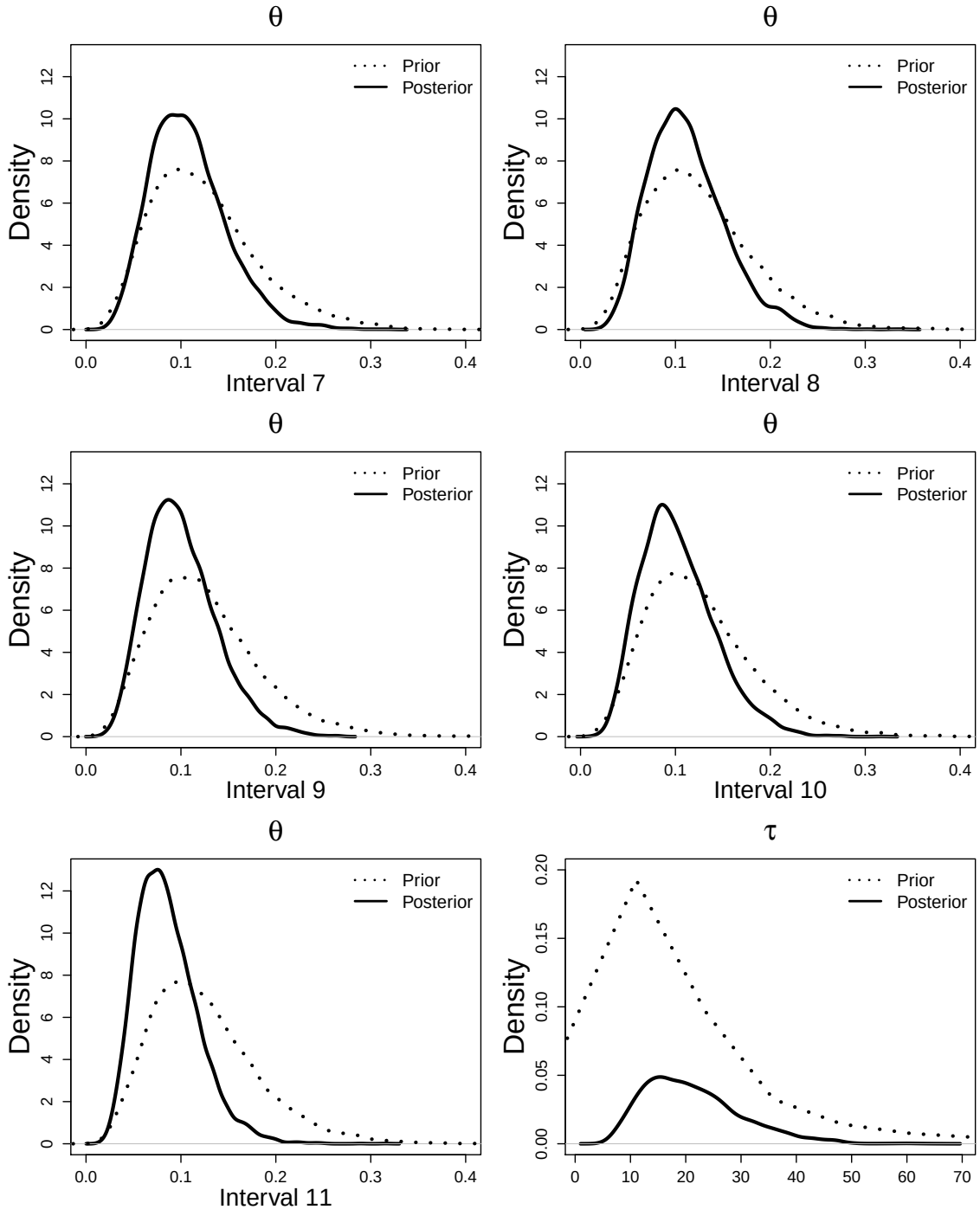

Figure 3.26: Prior and posterior distributions for the piecewise baseline hazard function for at-risk participants, θ, and hyperparameter τ (bottom-right panel) from the **country-stratified compound Poisson frailty** model analysis of **Partners PrEP** data.

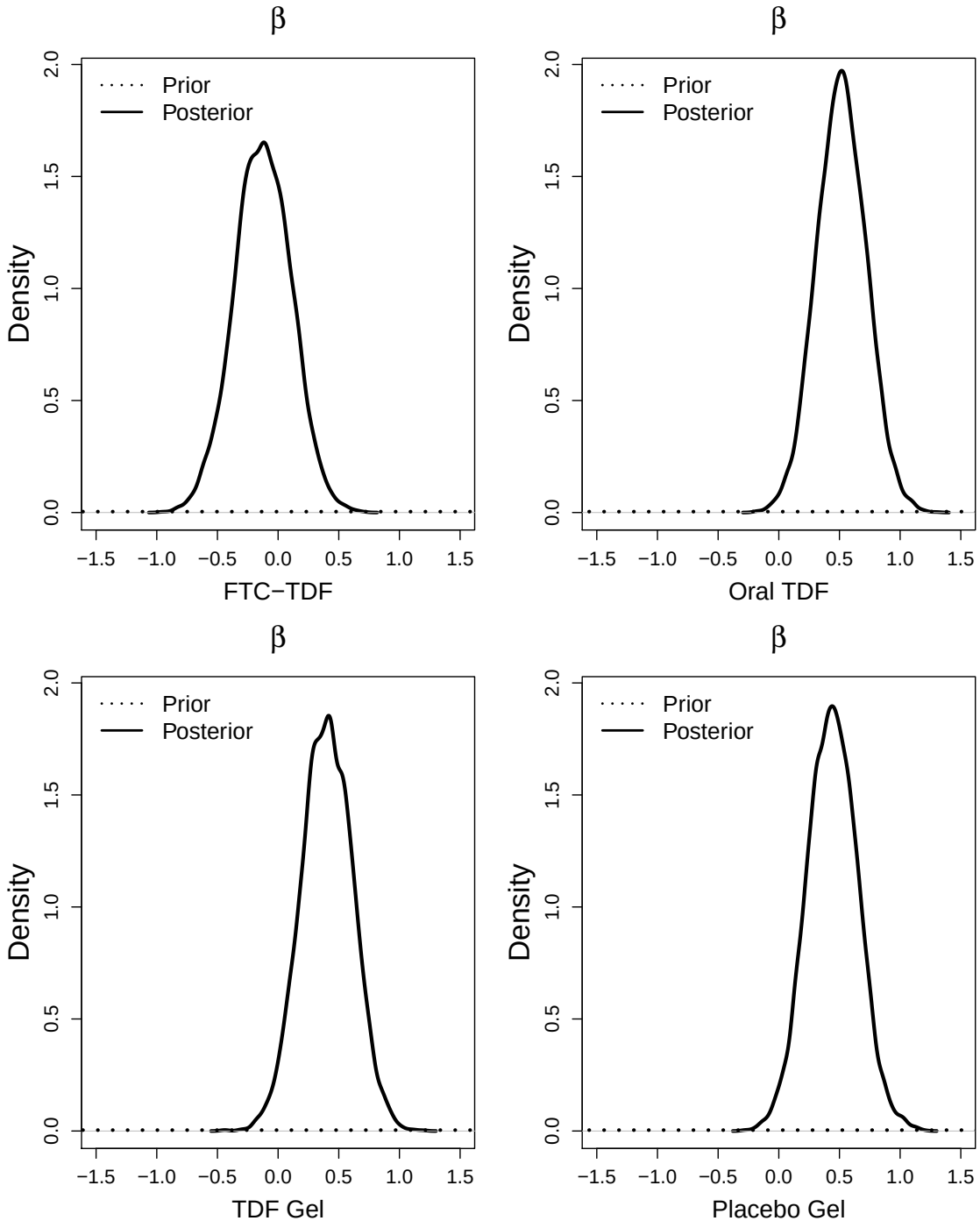

Figure 3.27: Prior and posterior distributions of log hazard ratio vector, $\boldsymbol{\beta}$, from the **latent class compound Poisson mixture frailty** model analysis of **VOICE** data.

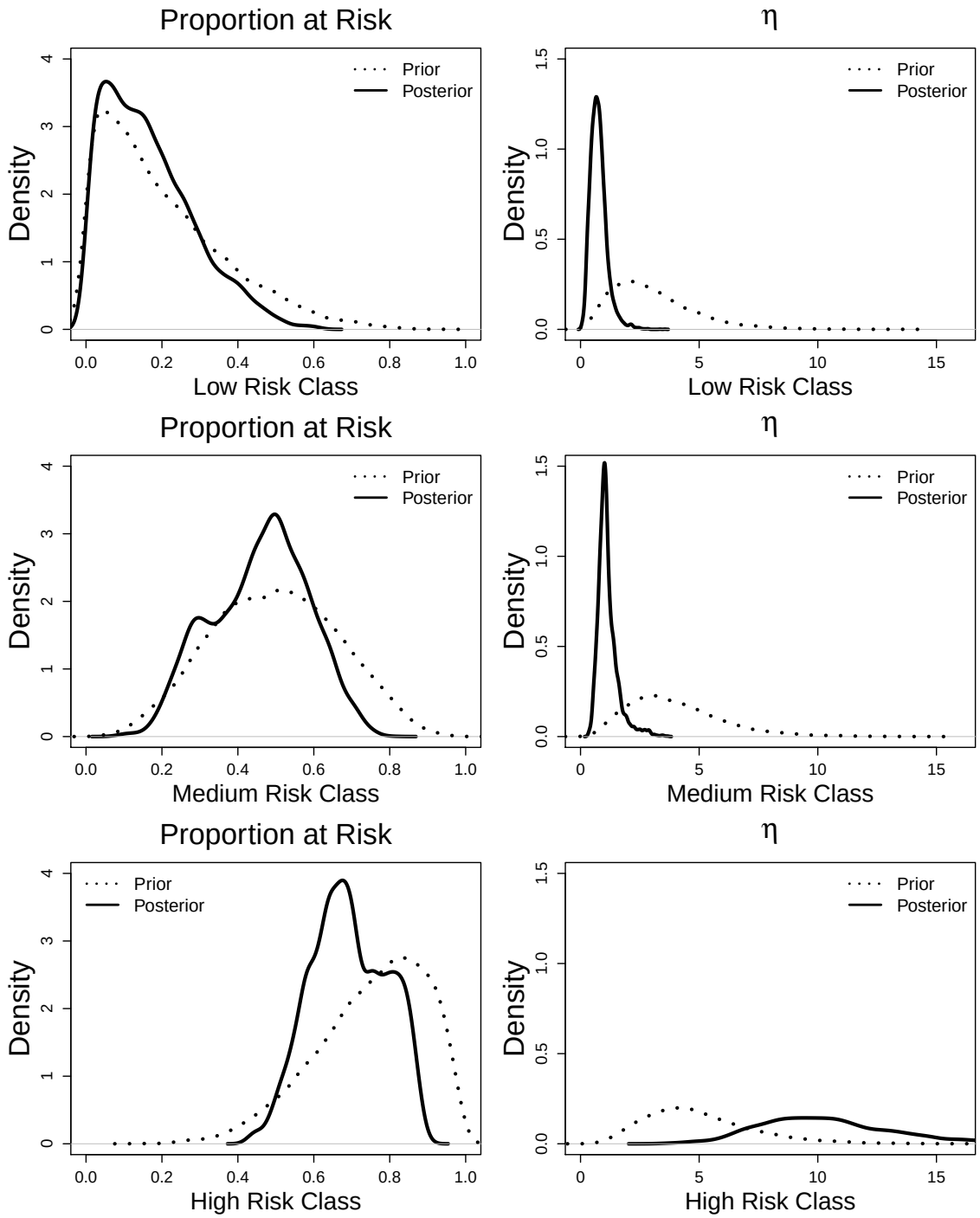

Figure 3.28: Prior and posterior distributions of the proportion at risk and η for each latent class from the **latent class compound Poisson mixture frailty** model analysis of **VOICE** data.

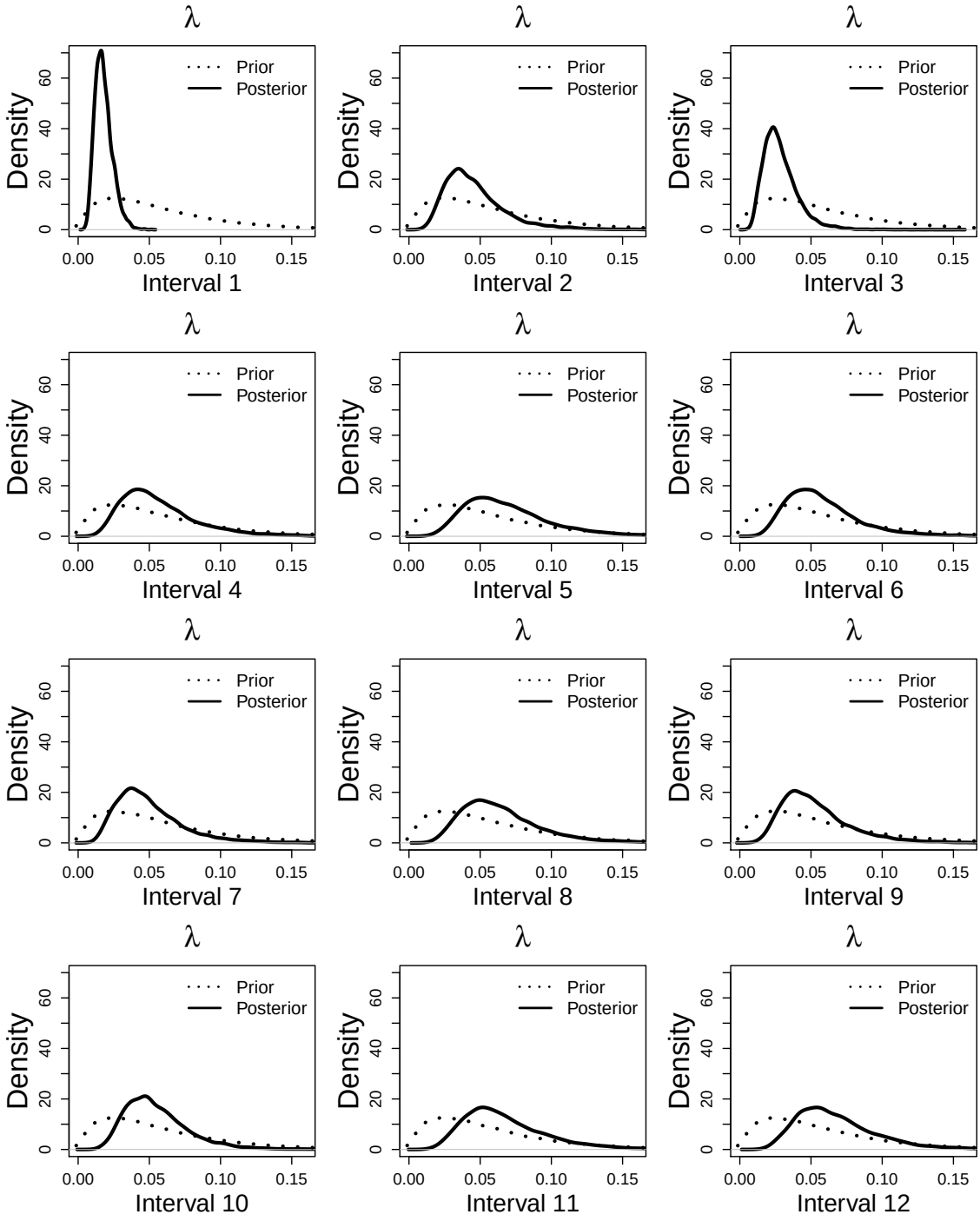

Figure 3.29: Prior and posterior distributions of piecewise constant baseline hazard from the **latent class compound Poisson mixture frailty** model analysis of **VOICE** data.

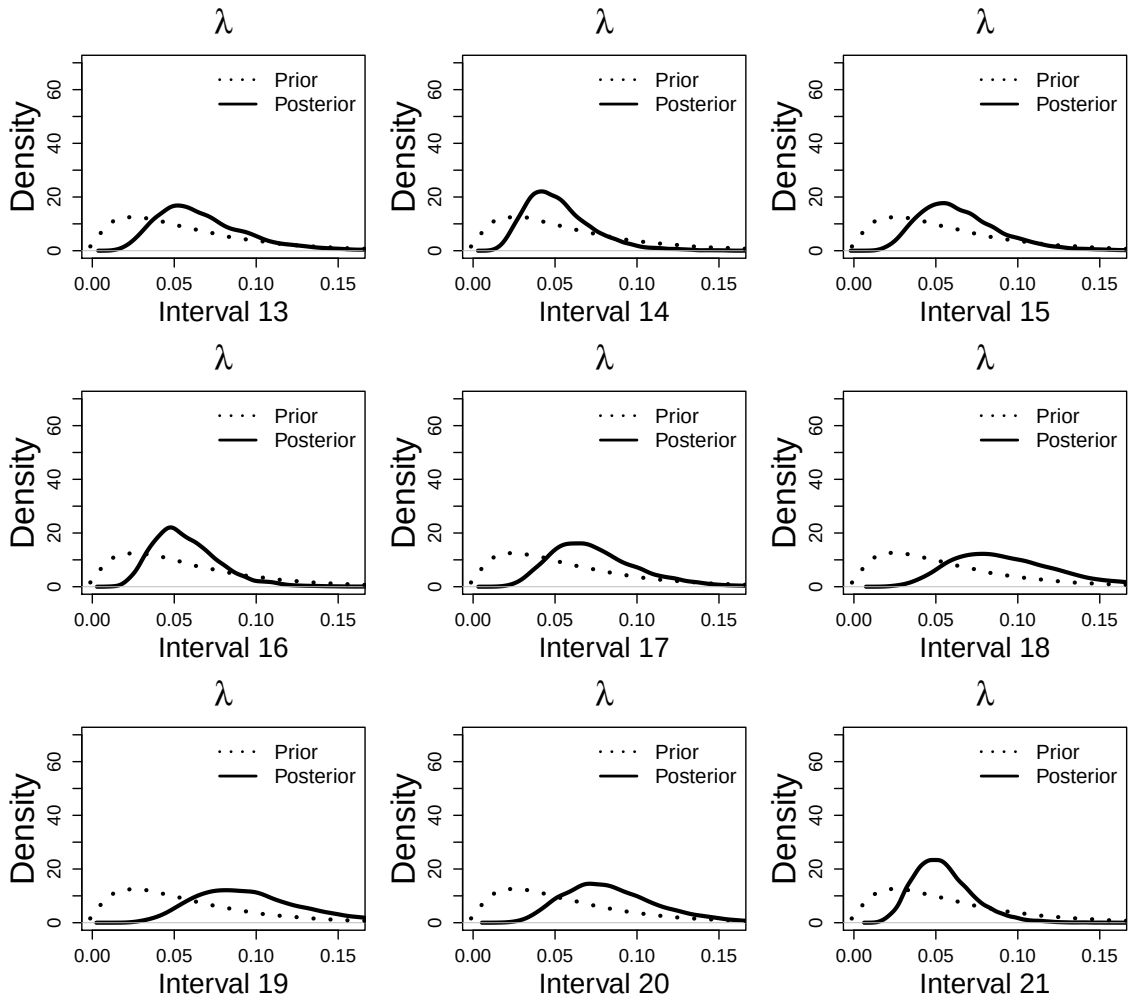

Figure 3.30: Prior and posterior distributions of piecewise constant baseline hazard from the **latent class compound Poisson mixture frailty** model analysis of **VOICE** data.

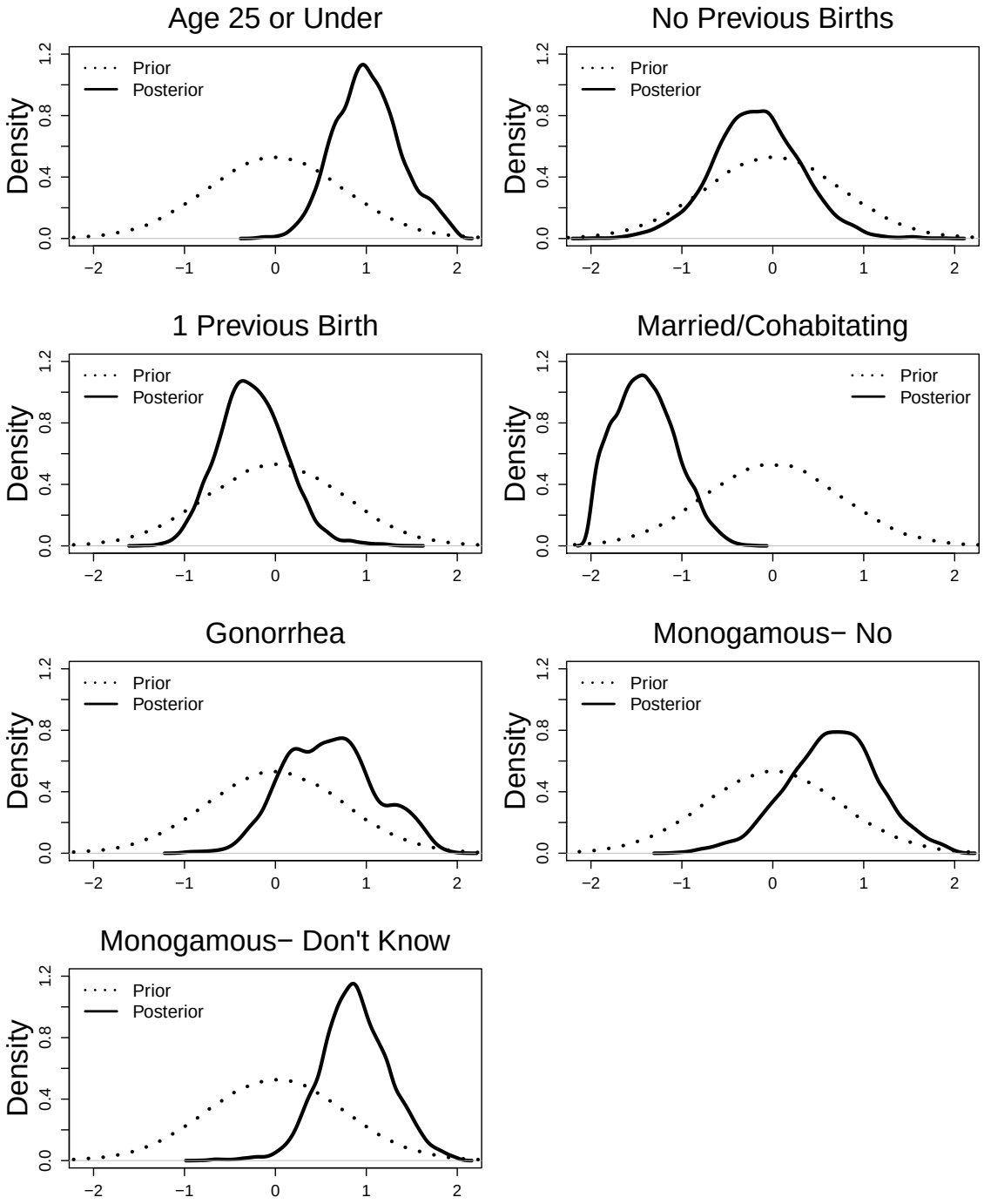

Figure 3.31: Prior and posterior distributions for proportional odds model covariate effects, $\boldsymbol{\gamma}^*$, from the **latent class compound Poisson mixture frailty** model analysis of **VOICE** data.

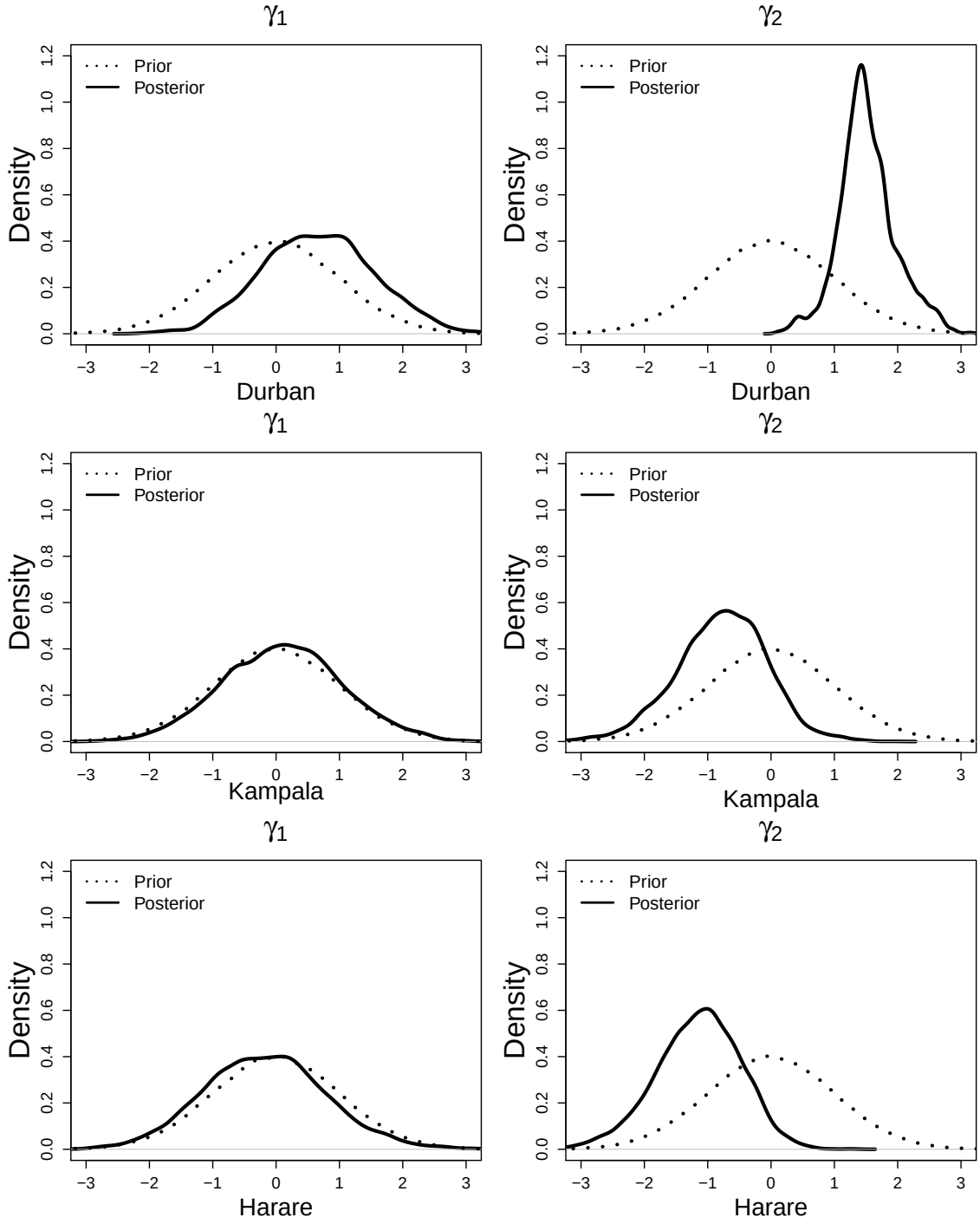

Figure 3.32: Prior and posterior distributions for non-proportional site location effects, $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, from the **latent class compound Poisson mixture frailty** model analysis of **VOICE** data.

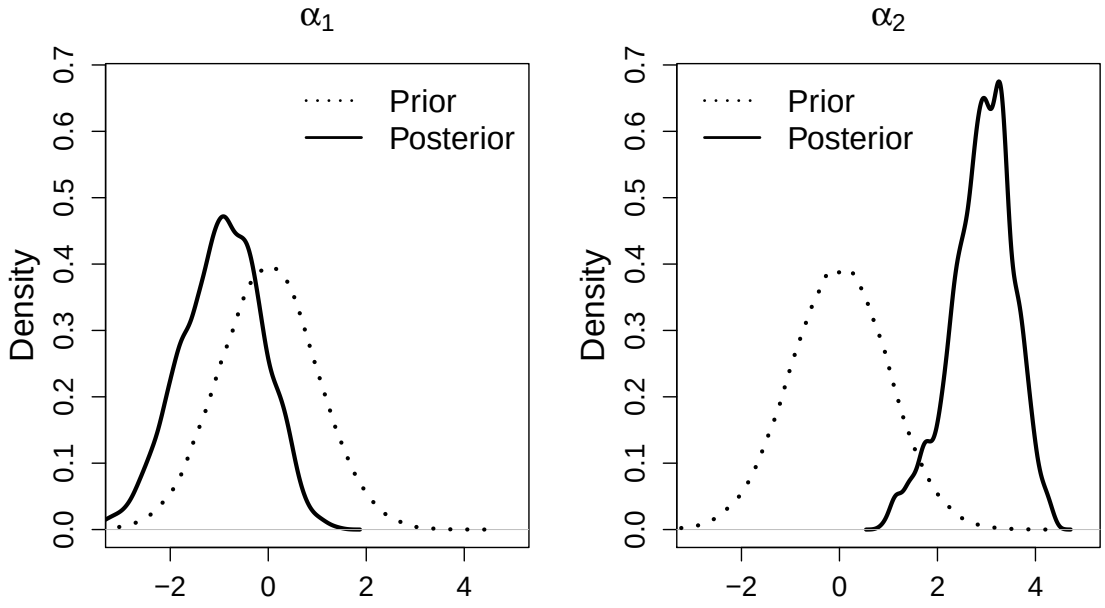

Figure 3.33: Prior and posterior distributions for proportional odds model intercepts α_1 and α_2 from the **latent class compound Poisson mixture frailty** model analysis of **VOICE** data.

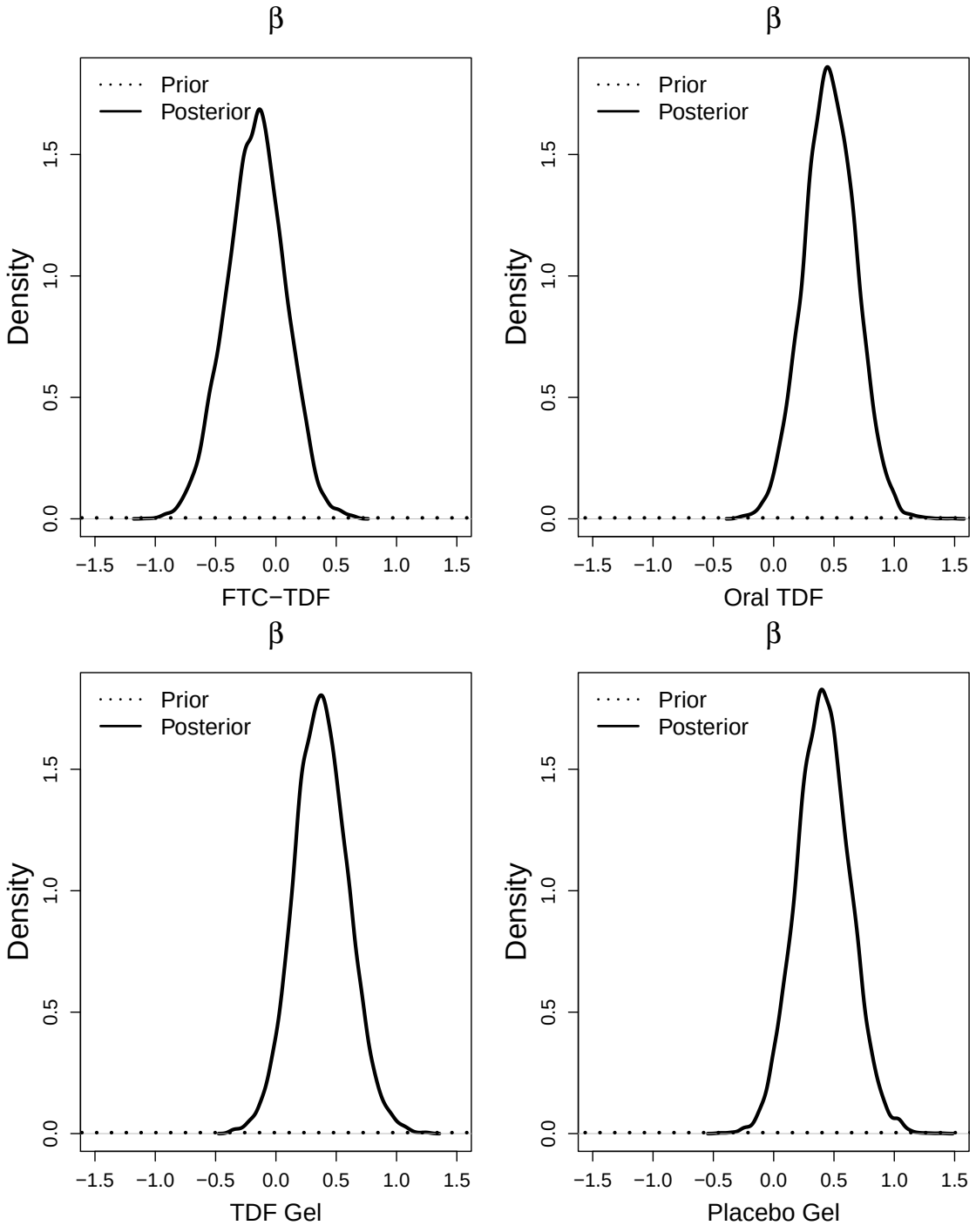

Figure 3.34: Prior and posterior distributions for log- hazard ratios from the **site location-stratified compound Poisson frailty** model analysis of **VOICE** data.

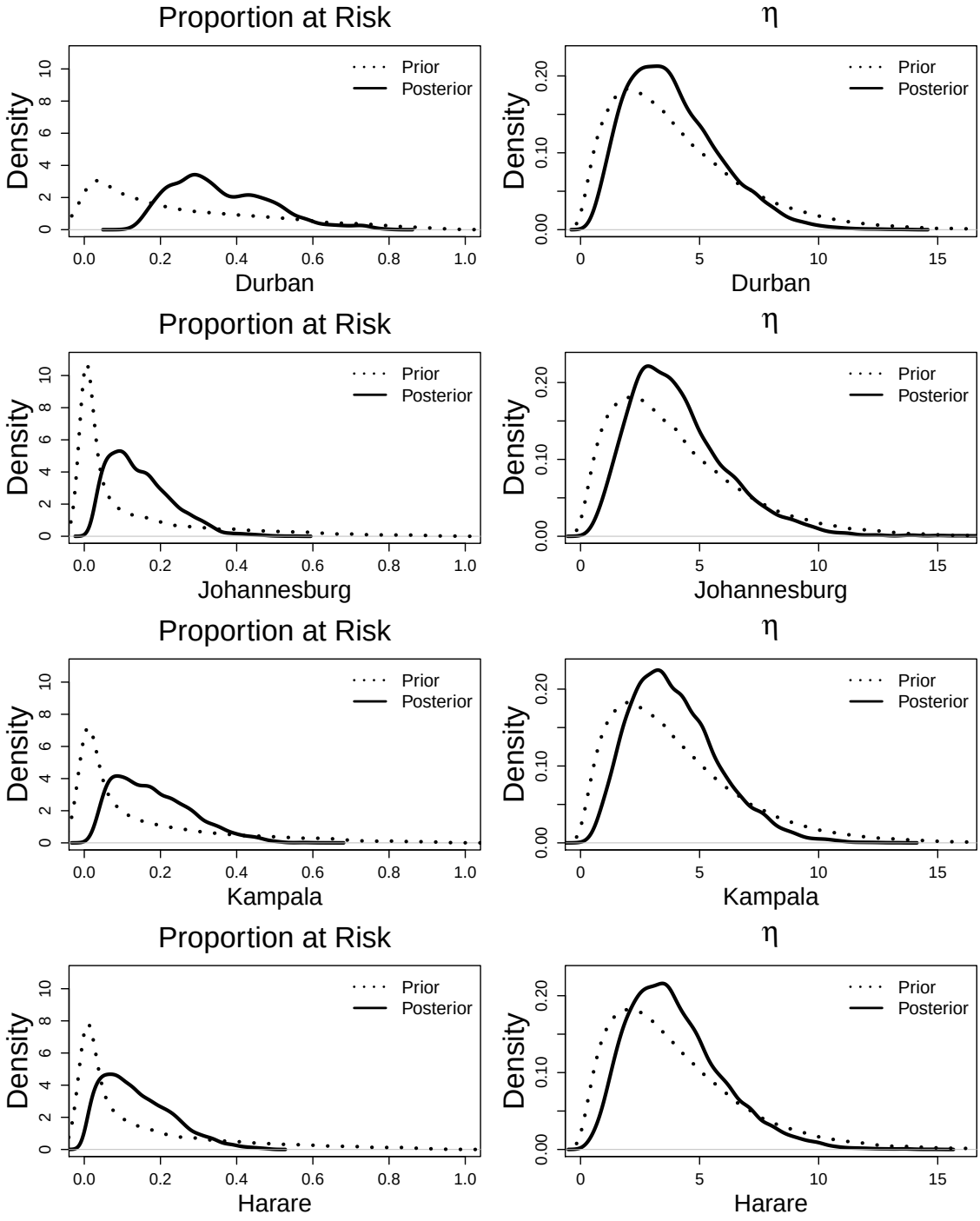

Figure 3.35: Prior and posterior distributions for proportion at risk and η from the **site location-stratified compound Poisson frailty** model analysis of **VOICE** data.

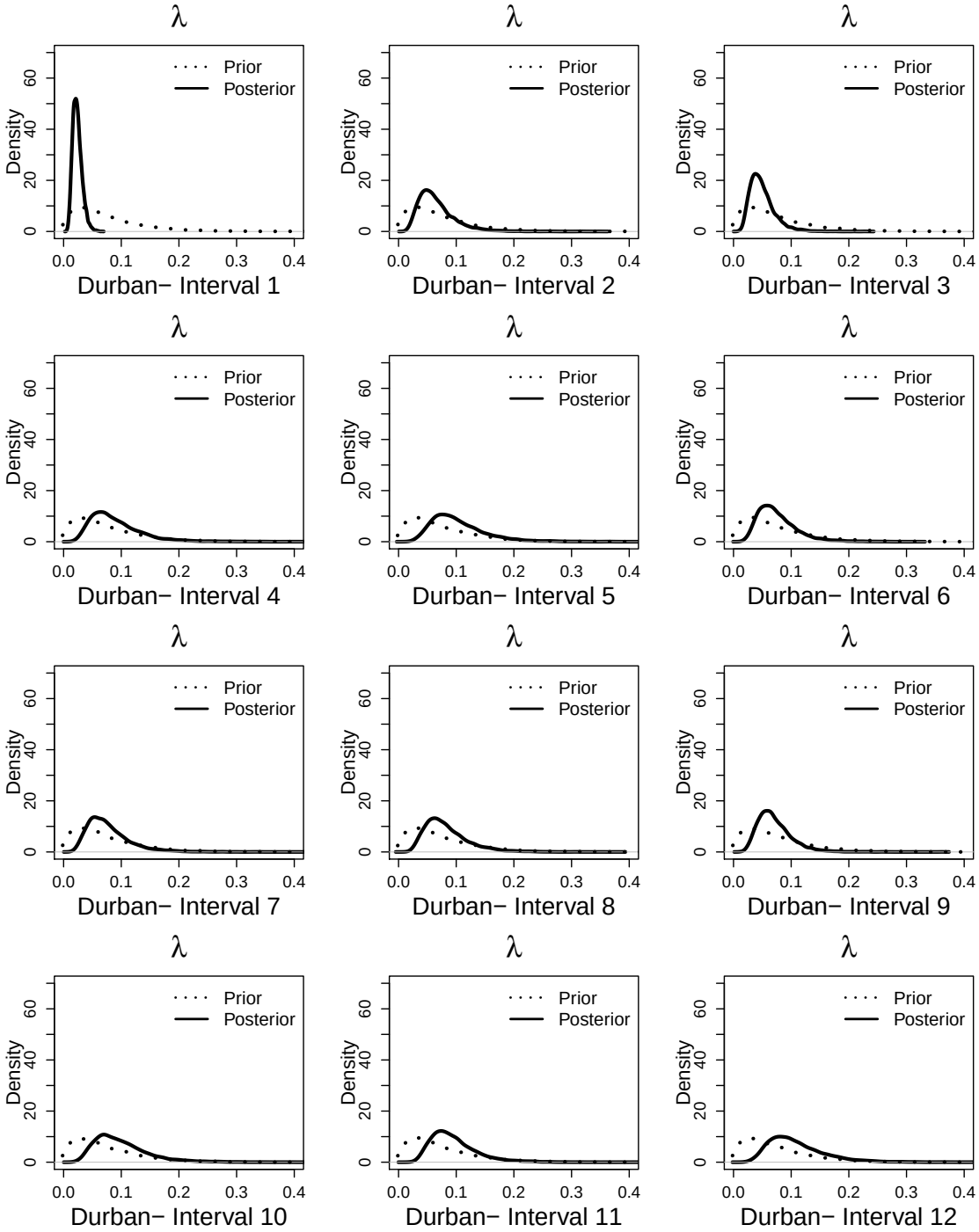

Figure 3.36: Prior and posterior distributions for piecewise baseline hazard function for sites around Durban, South Africa from the **site location-stratified compound Poisson frailty** model analysis of **VOICE** data.

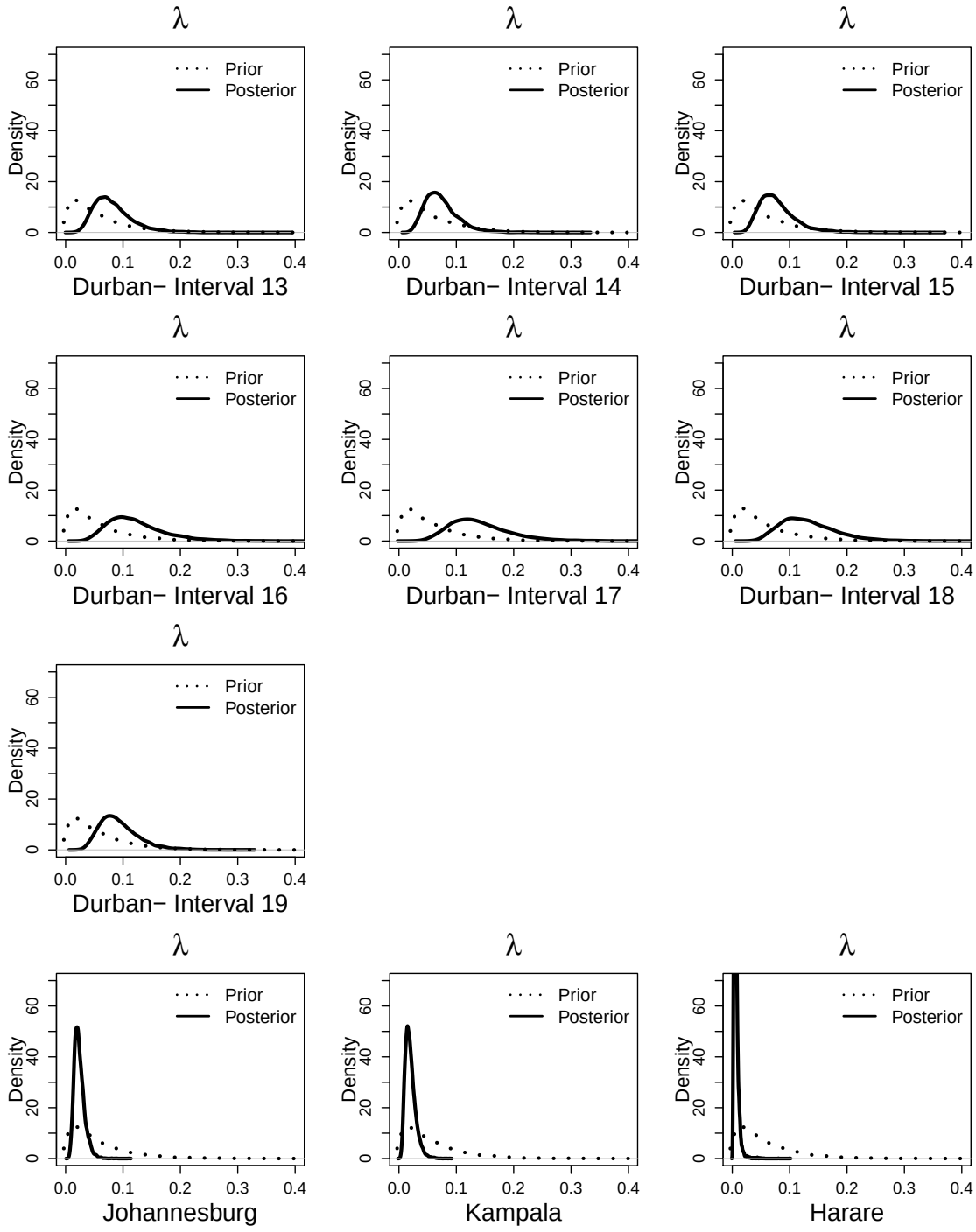

Figure 3.37: Prior and posterior distributions for piecewise baseline hazard function for sites around Durban, South Africa (upper 3 rows) and sites around Johannesburg, South Africa, Kampala, Uganda, and Harare, Zimbabwe (lower row) from the **site location-stratified compound Poisson frailty** model analysis of **VOICE** data.

Chapter 4

FUTURE WORK

In previous chapters, we have focused on the development of compound Poisson frailty models to adjust for heterogeneity in survival analysis. In this chapter, we first introduce ideas for a related model suitable when the number of exposure processes is observed. Then, we consider other areas of future research related to the compound Poisson frailty models proposed in Chapters 2 and 3.

4.1 Bayesian Hierarchical Gamma Frailty Model with Subject-specific Parameters

4.1.1 Introduction

In this section, we outline a frailty model to adjust for heterogeneity in survival analysis in a situation where (1) an individual's risk is the result of independent, competing exposure processes, (2) the number of exposure processes is observed for each individual, and (3) covariate data are also available on the exposure processes. Instead of assuming the level of risk associated with each of an individual's exposure processes follows the same distribution, as in the models proposed in Chapters 2 and 3, we introduce ideas here for a model that enables us to vary the frailty distribution based on characteristics observed at both the exposure process and the individual level.

As an example, consider the relationship between atherosclerotic lesions in coronary arteries and coronary heart disease (CHD). An atherosclerotic lesion can lead to the formation of a blood clot which blocks the artery, stopping the flow of blood and triggering a CHD-related event such as a myocardial infarction or cardiac arrest. In our framework, atherosclerotic lesions are exposure processes, each with the potential to cause a CHD event. The level of risk posed by each lesion varies depending on lesion-specific characteristics such as diameter, volume, and location. There are risk factors common to all lesions within

an individual as well, such as whether the individual has hypertension or diabetes or is a smoker. These individual-level risk factors also influence the risk posed by each lesion. The risk associated with each lesion depends on both observed lesion- and individual-specific characteristics as well as those which are not observed. Under the assumption that lesions act independently to trigger an event, an individual's total risk of a CHD event is the sum of risks from all lesions. Heterogeneity in risk of an event among individuals is the result of variability in the number of atherosclerotic lesions and the risk associated with each. We seek a survival model that reflects this relationship.

In contrast to our previous applications where risk processes are not observed, computed tomography (CT) enables us to observe the number of atherosclerotic lesions an individual has, as well as to assess characteristics of each lesion. Since the number of exposure processes is observed, it is not necessary to model it as an unobserved random variable. Furthermore, the risks associated with each lesion can be modeled separately, depending on individual- and lesion-level characteristics. Thus, we can consider frailty partially informed for individuals; the number of exposure processes is observed but the risk associated with each is not. Instead of modeling frailty on the level of the individual, we model it for each exposure process. In the proposed model, the shape parameter of the gamma distribution modeling frailty for each lesion is regressed on lesion- and individual-level covariates via a generalized linear model. Thus, the risk associated with each lesion follows a different distribution depending on these covariates. In this way, the model allows risk-related characteristics to inform an individual's risk but, recognizing some risk factors are not observed, still allows for random variability. Next, we define an individual's risk hierarchically: the total risk for an individual is the sum of the independent (but not identically distributed) frailties associated with each lesion. Said another way, frailty for an individual is the sum of that individual's exposure process-level frailties. As a result, frailty is modeled with a gamma distribution with subject-specific parameters.

While we have motivated this research from a competing risks perspective, the proposed model addresses another problem in survival analysis, that of clustered time-to-event data where the event is observed on the cluster (individual)-level while the unit (lesion) experiencing the event is unknown. In the language of our research and application, each

individual is a cluster and lesions are exposure processes clustered on the individual level. Although each lesion could lead to a CHD event, only the first event is typically of interest in survival studies, and the lesion which experienced the event is not observed.

There are two primary limitations to modeling risk in this setting: First, current frailty models for clustered data provide no mechanism for using any risk-related information available on the unit level. While many approaches exist for modeling heterogeneity in clustered survival data, these methods use only cluster-level information to model frailty; units within a cluster either share a frailty distribution or a frailty term [30, 38, 39]. The proposed model better reflects sources of heterogeneity in risk by incorporating unit-level covariate data.

Second, existing methods for analyzing clustered time-to-event data assume that an event time is observed for each unit within a cluster. When this is not the case, survival is typically modeled on the cluster level, enabling estimation of the effect of cluster-level covariates on hazard but either ignoring or aggregating unit-level covariates. For example, a recent study by Kuk et al (2014) examined the relationship between ulcer volume in the carotid arteries and vascular events. While researchers were able to measure the number of ulcers present as well as the volume of each, their analysis estimated the effect of total carotid ulcer volume on risk of an event.

In the context of atherosclerosis, CHD-free survival is modeled on the individual level, despite events occurring at the lesion level. Additionally, the effect of individual-level covariates, for example, smoking or diabetes, can be modeled, either in the regression term of a proportional hazards model or through a frailty distribution (as was proposed in Chapter 3), while lesion-level characteristics must be summarized across all lesions if they are to be included in the model. For example, the Agatston score, a measure of the amount of coronary artery calcium present, can be calculated for each lesion using CT, but the sum of all lesion-specific scores is usually included in individual-level survival models [6]. In the proposed model, time-to-event is modeled on the individual level but risk is modeled in such a way as to adjust for both individual- and lesion-level characteristics as well as unobserved heterogeneity.

We propose a Bayesian hierarchical gamma frailty model with subject-specific param-

eters to account for clustered time-to-event data where events are observed at the cluster (individual) level and the unit (exposure process) experiencing the event is unknown. The proposed model reflects underlying biological risk mechanisms in this setting better than existing approaches by modeling risk at the level of the exposure process. Exposure process-specific frailty distributions allow both exposure process- and individual-level covariates to influence risk while also accounting for unobserved heterogeneity.

This section proceeds as follows. In Section 4.1.2, we present a hierarchical frailty model in which parameters of the gamma distribution modeling the risk associated with each exposure process vary according to a generalized linear regression of exposure process- and individual-level covariates. Bayesian estimation procedures for this model are also outlined. In Section 4.1.3, we detail a planned application to atherosclerosis and coronary heart disease. We close with a discussion.

4.1.2 Methods

We begin with a definition of a hierarchical gamma frailty model that models risk of an event using both individual- and exposure process-level covariates before discussing prior elicitation and Bayesian estimation procedures for the model.

Hierarchical Gamma Frailty Model with Subject-specific Parameters

Our hierarchical model for risk is motivated by the compound Poisson frailty distribution, where an individual's risk is the result of an unknown number of independent, competing exposure processes. In this situation, the number of exposure processes is known. We model the risk associated with each exposure process using a distribution that varies depending on both individual- and exposure process-level covariates. An individual's risk is then the sum of risks across exposure processes.

We begin by introducing a model for the risk of event associated with unique exposure processes that reflects underlying biological risk mechanisms. Let N_i be the observed number of exposure processes for individual i, $i = 1, \ldots, n$. We can define the hazard for the ith

participant's mth exposure process as $h(t) = \lambda(t)U_{im}$ where

$$U_{im} \sim Gamma(\eta_{im}, \nu) \qquad \text{for } m = 1, \ldots, N_i,$$

$\lambda(t)$ is a baseline hazard function common to all exposure processes, and $Gamma(\eta, \nu)$ is the gamma distribution with shape η, rate ν, and expectation $E(U_{im}) = \eta_{im}/\nu$. Next, the risk of an event may vary across exposure processes according to measurable characteristics. We accommodate this by allowing the parameter η_{im} to vary across exposure processes via a generalized linear regression on a Q-length vector of covariates, $\mathbf{v}_{im}$:

$$g(\eta_{im}|\mathbf{v}_{im}) = \mathbf{v}'_{im}\boldsymbol{\xi} \tag{4.1}$$

where $\boldsymbol{\xi}$ is a Q-length parameter vector common to all exposure processes that relates covariates, $\mathbf{v}_{im}$, to the shape parameter, η_{im}. By modeling the effect of covariates on a frailty model parameter, we allow these covariates to increase or decrease risk in a stochastic manner. Risk-related covariates determine the distribution of frailty for each exposure process, but they do not determine frailty. As a result, exposure processes with more risk factors have, on average, a higher frailty than those with fewer risk factors. However, it is also possible for an exposure process with fewer risk factors to have frailty greater than an exposure process with more risk factors. In doing so, this framework also allows for unobserved heterogeneity.

We first consider the link function g and covariate vector $\mathbf{v}_i$ from Equation (4.1) before aggregating exposure process-specific frailties on the individual level. First, the log function is an obvious choice for link g in Equation (4.1) as it bounds the response above zero, but other link functions may also be appropriate depending on the application.

Next, the covariate vector $\mathbf{v}_{im}$ may include both individual and exposure process-specific covariates. All risk-related exposure process characteristics should be included in $\mathbf{v}_{im}$ because the individual-level survival model (defined below in Section 4.1.2) does not accommodate exposure process-level covariates, as discussed in Section 4.1.1. Individual-level covariates included in $\mathbf{v}_{im}$ will depend on the application but should include factors that are associated with the level of risk posed by an exposure process and exclude factors for which a hazard ratio estimate is sought. Effects are not guaranteed to be identifiable for

covariates included in both the exposure process-level regression model for η_{im} given in Equation (4.1) and the individual-level proportional hazards function. Preliminary simulations suggest that the proposed model is not able to distinguish between a covariate's indirect effects on the risk associated with each exposure process and its direct effects on the individual's hazard. In particular, all effects are attributed to the frailty model while coefficients in the proportional hazards function are estimated to be zero.

We continue by defining an individual's frailty hierarchically. While frailty is modeled separately for each exposure process, it must be aggregated on the individual level to be included in the survival model because, as discussed earlier in Section 4.1.1, survival must be modeled on the level at which the event is observed. Assuming that an individual's exposure processes pose independent, competing risks, these risks are additive across exposure processes. Thus, a participant's risk of an event at time t is the sum of risks associated with each exposure process: $\lambda(t)\sum_{m=1}^{N_i} U_{im}$. We define Z_i, the individual-level frailty for participant i, as follows:

$$Z_i = \sum_{m=1}^{N_i} U_{im} \sim Gamma\Big(\sum_{m=1}^{N_i} \eta_{im}, \nu\Big).$$

where $\sum_{m=1}^{N_i} \eta_{im}$ is the shape parameter specific to individual i. If the number of exposure processes for an individual equals zero, then frailty necessarily equals zero, i.e. no risk. In comparison to frailty models proposed in Chapters 2 and 3, the frailty for individual i, Z_i, can be considered partially informed because the number of exposure processes is observed but the risk associated with each is not.

As in previously proposed models, it is necessary for model identifiability that the expected frailty for a randomly selected individual equals one. In this model, the expectation of Z_i for a particular individual i is $\sum_{m=1}^{N_i} \eta_{im} \,/\, \nu$. For the expectation of a randomly selected Z_i to equal one, we set ν equal to the mean of $\sum_{m=1}^{N_i} \eta_{im}$, the shape parameter for the individual-level frailty distribution, over all individuals:

$$\begin{aligned} \nu &= \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{N_i} \eta_{im} \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{N_i} g^{-1}(\mathbf{v}'_{im}\boldsymbol{\xi}) \end{aligned} \tag{4.2}$$

where the last equality follows from Equation (4.1).

We conclude our definition of this frailty distribution by examining how defining the rate parameter ν as a function of all covariates $\mathbf{v}_{im}$ affects the interpretation of $\boldsymbol{\xi}$. Under this restriction, the expected frailty of exposure process m for participant i with predictor v_{im}, U_{im}, is:

$$E(U_{im}|v_{im}) = \frac{g^{-1}(v_{im}\xi)}{\frac{1}{n}\Big(\sum_{i'\neq i}\sum_{m=1}^{N_{i'}} g^{-1}(v_{i'm}\xi) + \sum_{m'\neq m} g^{-1}(v_{im'}\xi) + g^{-1}(v_{im}\xi)\Big)}$$

where the numerator is equal to η_{im}, the shape parameter for the ith participant's mth exposure process, and the denominator is equal to ν, the average of the sum of all exposure process-level shape parameters across all participants. (The first term after the factor of $\frac{1}{n}$ in the denominator is the sum of shape parameters for all exposure processes for individuals other than i, the second term is the sum of shape parameters for all exposure processes for individual i except m, and the third term is the shape parameter for the ith individual's mth exposure process.) If exposure process m for participant i instead has predictor $v_{im}+1$, both the shape parameter for that exposure process's frailty distribution, η_{im}, and the rate, ν, change:

$$E(U_{im}|v_{im}+1) = \frac{g^{-1}(v_{im}\xi+\xi)}{\frac{1}{n}\Big(\sum_{i'\neq i}\sum_{m=1}^{N_{i'}} g^{-1}(v_{i'm}\xi) + \sum_{m'\neq m} g^{-1}(v_{im'}\xi) + g^{-1}(v_{im}\xi+\xi)\Big)}.$$

Assuming a log link for g, the expected frailty given v_{im} and $v_{im}+1$, respectively, are:

$$\begin{aligned} E(U_{im}|v_{im}) &= \frac{\exp(v_{im}\xi)}{\frac{1}{n}\Big(\sum_{i'\neq i}\sum_{m=1}^{N_{i'}} \exp(v_{i'm}\xi) + \sum_{m'\neq m} \exp(v_{im'}\xi) + \exp(v_{im}\xi)\Big)} \\ E(U_{im}|v_{im}+1) &= \frac{\exp(v_{im}\xi+\xi)}{\frac{1}{n}\Big(\sum_{i'\neq i}\sum_{m=1}^{N_{i'}} \exp(v_{i'm}\xi) + \sum_{m'\neq m} \exp(v_{im'}\xi) + \exp(v_{im}\xi+\xi)\Big)}. \end{aligned}$$

It follows that the change in expected value for an exposure process's frailty associated with a unit increase in covariate v_{im} is:

$$\frac{E(U_{im}|v_{im}+1)}{E(U_{im}|v_{im})} = \exp(\xi)\,\frac{\sum_{i'\neq i}\sum_{m=1}^{N_i'} \exp(v_{i'm}\xi) + \sum_{m'\neq m} \exp(v_{im'}\xi) + \exp(v_{im}\xi)}{\sum_{i'\neq i}\sum_{m=1}^{N_i'} \exp(v_{i'm}\xi) + \sum_{m'\neq m} \exp(v_{im'}\xi) + \exp(v_{im}\xi+\xi)}. \quad (4.3)$$

We see here that an interpretation for ξ is not clear when using a log link, at least in small sample sizes. However, in a setting with a large number of individuals, a large number of

exposure processes per individual, or both, a single exposure process will have very little effect on ν. We find the limit of the righthand side of Equation (4.3) to be

$$\lim_{n,m\to\infty} \exp(\xi)\,\frac{\sum_{i'\neq i}\sum_{m=1}^{N_i'}\exp(v_{i'm}\xi)+\sum_{m'\neq m}\exp(v_{im'}\xi)+\exp(v_{im}\xi)}{\sum_{i'\neq i}\sum_{m=1}^{N_i'}\exp(v_{i'm}\xi)+\sum_{m'\neq m}\exp(v_{im'}\xi)+\exp(v_{im}\xi+\xi)} = \exp(\xi)$$

and conclude that $\exp(\xi)$ can be interpreted as the change in the expected frailty for an exposure process associated with a unit increase in covariate v when the sample sizes or number of exposure processes per individual is large.

Hazard Regression with Frailty

We next define a survival model with frailty on the individual level. For each participant i, Z_i, the sum of all exposure process-specific frailties, $U_{i1},\ldots,U_{iN_i}$, is included in the individual's hazard function as a frailty term which modifies the hazard as appropriate:

$$\begin{aligned}\lambda_i(t|\mathbf{x}_i) &= \lambda(t)\sum_{m=1}^{N_i} U_{im}\exp(\mathbf{x}_i'\boldsymbol{\beta}) \\ &= \lambda(t)\,Z_i\exp(\mathbf{x}_i'\boldsymbol{\beta}) \qquad (4.4)\end{aligned}$$

where $\exp(\mathbf{x}_i'\boldsymbol{\beta})$ is a proportional hazards function modeling the effects of individual-level covariates $\mathbf{x}_i$ on risk. As mentioned in Section 4.1.2, $\mathbf{x}_i$ should include individual-level covariates for which a hazard ratio estimate is sought.

Following from Equation (4.4), the individual-specific density and survival functions are:

$$\begin{aligned}f(t|\mathbf{x}_i,Z_i) &= \lambda(t)\,Z_i\exp(\mathbf{x}_i'\boldsymbol{\beta})\,S(t|x_i,Z_i) \\ S(t|\mathbf{x}_i,Z_i) &= \exp\Big(-Z_i\exp(\mathbf{x}_i'\boldsymbol{\beta})\,\Lambda(t)\Big),\end{aligned}$$

where $\Lambda(t)=\int_0^t\lambda(t)\,dt$ is the cumulative baseline hazard at time t.

Note that inclusion of frailty in the Cox model changes the interpretation of each parameter in vector $\boldsymbol{\beta}$ from a population-averaged log-hazards ratio to the log-hazards ratio for an individual. We contrast the interpretation of the HR with that obtained in the compound Poisson frailty models proposed in Chapters 2 and 3 where $\exp(\beta)$ is the individual-level hazard ratio for those at risk. In this model, $\exp(\beta)$ is also interpreted as the individual-level hazard ratio for those at risk with the key distinction being that, in this setting, we observe which individuals are at risk, i.e. have a positive number of exposure processes.

Model Likelihood

Following the definition of a hierarchical gamma frailty distribution with subject-specific shape parameter and survival model, we can define the model likelihood given time-to-event data. We observe (t_i, δ_i) where t_i is potentially censored time-to-event and δ_i is an indicator variable equaling one if an event was observed and zero otherwise for individual i, $i = 1, \ldots, n$. The likelihood for this model can be written as:

$$
\begin{aligned}
&L\big(\boldsymbol{\beta}, \lambda(t), \boldsymbol{\xi} | (t_i, \delta_i, \mathbf{x}_i, \underline{\mathbf{v}_i}, N_i, Z_i), i = 1, \ldots, n\big) \\
&\quad = \prod_{i=1}^{n} f(Z_i | N_i, \underline{\mathbf{v}_i}, \boldsymbol{\xi})\, f(t_i, \delta_i | \mathbf{x}_i, Z_i, \boldsymbol{\beta}, \lambda(t)) \\
&\quad = \prod_{i=1}^{n} \Big[f(Z_i | N_i, \underline{\mathbf{v}_i}, \boldsymbol{\xi})\, f(t_i, \delta_i | \mathbf{x}_i, Z_i, \boldsymbol{\beta}, \lambda(t)) \Big]^{\mathbf{1}_{[N_i > 0]}} \\
&\qquad\qquad \times \Big[f(Z_i | N_i, \underline{\mathbf{v}_i}, \boldsymbol{\xi})\, f(t_i, \delta_i | \mathbf{x}_i, Z_i, \boldsymbol{\beta}, \lambda(t)) \Big]^{\mathbf{1}_{[N_i = 0]}} \\
&\quad = \prod_{i=1}^{n} \Big[\frac{\nu^{\sum_{m=1}^{N_i} \eta_{im}}}{\Gamma(\sum_{m=1}^{N_i} \eta_{im})}\, Z_i^{\sum_{m=1}^{N_i} \eta_{im} - 1}\, e^{-Z_i \nu}\, (\lambda(t) Z_i e^{\mathbf{x}_i \boldsymbol{\beta}})^{\delta_i}\, e^{-Z_i e^{\mathbf{x}_i \boldsymbol{\beta}} \Lambda(t_i)} \Big]^{\mathbf{1}_{[N_i > 0]}} \\
&\qquad\qquad \times \Big[\Delta_0(Z_i)\, \Delta_\infty(t_i) \Big]^{\mathbf{1}_{N_i = 0}} \\
&\quad = \prod_{i=1}^{n} \Big[\frac{\nu^{\sum_{m=1}^{N_i} g^{-1}(\mathbf{v}'_{im} \boldsymbol{\xi})}}{\Gamma(\sum_{m=1}^{N_i} g^{-1}(\mathbf{v}'_{im} \boldsymbol{\xi}))}\, Z_i^{\sum_{m=1}^{N_i} g^{-1}(\mathbf{v}'_{im} \boldsymbol{\xi}) - 1}\, e^{-Z_i \nu}\, (\lambda(t) Z_i e^{\mathbf{x}_i \boldsymbol{\beta}})^{\delta_i}\, e^{-Z_i e^{\mathbf{x}_i \boldsymbol{\beta}} \Lambda(t_i)} \Big]^{\mathbf{1}_{[N_i > 0]}} \\
&\qquad\qquad \times \Big[\boldsymbol{\Delta_0}(Z_i)\, \boldsymbol{\Delta_\infty}(t_i) \Big]^{\mathbf{1}_{N_i = 0}}
\end{aligned}
\tag{4.5}
$$

where $\underline{\mathbf{v}_i}$ is the matrix of covariate vectors $\big(\mathbf{v}_{i1}, \ldots, \mathbf{v}_{iN_i}\big)$ for participant i, Z_i is the sum of exposure-process specific frailties for participant i, $\sum_{m=1}^{N_i} U_{im}$ for $i = 1, \ldots, n$, and $\boldsymbol{\Delta_0}(Z_i)$ and $\boldsymbol{\Delta_\infty}(t_i)$ indicate a point mass for frailty, Z_i, at zero and time-to-event, t_i, at infinity for participant i with $N_i = 0$, i.e. no exposure processes. ν is equal to the mean shape parameter across all individuals, as defined in Equation (4.2), but dependency on all $\underline{\mathbf{v}_{i'}}$, $i' \neq i$, is suppressed in the likelihood above for clarity in presentation.

Bayesian Model Estimation

We complete Bayesian specification of the proposed model by considering prior elicitation, defining the joint posterior distribution, and discussing sampling methods for model parameters and latent variables.

Prior selection for model parameters will be application-specific. Priors for the baseline hazard function, $\lambda(t)$, and proportional hazards model coefficients, $\boldsymbol{\beta}$, can be similar to priors typically used for survival models [27]. If using a log-link for the generalized linear regression model in Equation (4.1), a multivariate normal prior on $\boldsymbol{\xi}$ would be convenient and easily accommodate prior belief about covariates' effects on the magnitude of risk for each unit.

After specifying priors on model parameters, the joint posterior distribution can be defined:

$$\begin{aligned} &\mathbf{p}\Big(\boldsymbol{\beta}, \lambda(t), \boldsymbol{\xi}, \mathbf{Z} | \big((t_i, \delta_i, \mathbf{x}_i, N_i, \underline{\mathbf{v}_i}), i = 1, \ldots, n\big), \boldsymbol{\Theta}\Big) \\ &\quad \propto L\big(\boldsymbol{\beta}, \lambda(t), \boldsymbol{\xi} | (t_i, \delta_i, \mathbf{x}_i, N_i, \underline{\mathbf{v}_i}, Z_i), i = 1, \ldots, n\big) \times \boldsymbol{\pi}\big(\boldsymbol{\beta}, \lambda(t), \boldsymbol{\xi} | \boldsymbol{\Theta}\big) \end{aligned} \tag{4.6}$$

where $\mathbf{p}(\cdot|\text{data})$ denotes the joint posterior density, $\boldsymbol{\pi}(\cdot|\boldsymbol{\Theta})$ denotes the joint prior for model parameters with hyperparameters $\boldsymbol{\Theta}$, and $\mathbf{Z}$ is the vector of individual-level frailty terms $(Z_1, Z_2, \ldots, Z_n)$. Conditional posteriors for each parameter follow from Equation (4.6) and are given in Appendix D. A Gibbs sampling algorithm is used to obtain samples from the joint posterior [21]. Slice sampling is used to sample from model parameters' posterior distributions when conjugate priors are not available [40]. The partially informed individual-level frailty, Z_i, is also sampled for all individuals, $i = 1, \ldots, n$, at each iteration of the Gibbs sampler. The conditional density for each Z_i is a gamma distribution with shape parameter equal to $\delta_i + \sum_{i=1}^{N_i} \eta_{im}$ and rate parameter equal to $\nu + \exp(\mathbf{x}_i'\boldsymbol{\beta})\Lambda(t_i)$.

4.1.3 Application

Data

To illustrate the proposed method, we will examine data from the Multi-Ethnic Study of Atherosclerosis (MESA). The MESA is a prospective cohort study with the objective of identifying characteristics related to progression of subclinical cardiovascular disease (CVD) [11]. Beginning in July 2000, the study enrolled over 6,500 men and women aged 45-84 years old and free of cardiovascular disease at enrollment from six study sites across the United States. Ethnic groups under-represented in existing literature on subclinical CVD

were targeted for enrollment in the MESA. The ethnic distribution of enrolled participants was 38% White, 28% African-American, 22% Hispanic, and 12% Chinese-American. Study design, data collection, and study outcomes are detailed in Bild et al (2002).

Chest CT was performed on all participants at the time of study entry. Details on scanning protocols and procedures for interpreting images are given in Carr et al (2005). The Agatston method was used to quantify the amount of coronary artery calcium for each lesion. Other lesion-level covariate information obtained from CT images include: artery location (left main, left anterior descending, left circumflex, right main), distance from artery ostium (as defined in Brown et al 2008), maximal diameter, and volume.

Model and Prior Specification

In our analysis, we will use data available on 3,128 participants with at least one calcified lesion detected by CT to model the time to *incident hard* CHD event, defined as definite or probable myocardial infarction, resuscitated cardiac arrest, or fatal CHD. Participants with no lesions detected are not included in the analysis; an individual with no lesions, according to this model, necessarily has no risk. For participants with multiple CT scans, one scan will be randomly selected to use in the analysis.

Since obtaining more information about subclinical CVD in underrepresented ethnic groups was a goal of the study, we will estimate the relative hazard of incident hard CHD events for African-American, Hispanic, and Chinese-American individuals in comparison to White individuals with the proposed model. It follows that $\mathbf{x}_i$ will be a covariate vector indicating race or ethnic group for each participant i, $i = 1, \ldots, n$. Diffuse normal priors will be specified for $\boldsymbol{\beta}$, the vector of log-hazard ratios associated with each group. For comparison, population-level hazard ratios for each race or ethnic group will also be estimated with the Cox proportional-hazards model.

Lesion- and participant- level covariates that are significant predictors of an event in a univariate logistic model will be used to model the shape parameter for the gamma distribution modeling the level of risk associated with each exposure process, i.e. $\mathbf{v}_{im}$ in Equation (4.1). These covariates include, on the participant-level: age, gender, body-mass

index (BMI), concentration of low and high density lipoproteins, and current smoking, and, on the lesion-level: Agatston score and volume. Normal priors will be placed on each component of $\boldsymbol{\xi}$, the effect of each lesion m-level covariate, v_{im}, on the shape parameter η_{im}, for all $m = 1, \ldots, N_i$, $i = 1, \ldots, n$.

We will allow for flexible estimation of the baseline hazard by assuming a piecewise constant function for $\lambda(t)$:

$$\lambda(t) = \lambda_s, \qquad t_s \leq t < t_{s+1},\ s = 1, \ldots, S$$

where t_s, $s = 1, \ldots, S$, denote cut-points, $t_1 = 0$, and $t_{S+1} = \infty$. Cut-points will be determined by events with approximately 8-10 events per interval. A gamma prior will be used for each λ_s as this is a conjugate prior.

Prior specification for model parameters will be done in such a way that expected incidence of CHD events among White participants is comparable to that reported for the United States (see, for example, [23]) while allowing for ample variability.

Posterior estimates for the proposed model will be obtained with a Gibbs sampling algorithm, as described in Section 4.1.2. Posterior medians and 95% highest posterior density intervals will be estimated from posterior samples for each model parameter (after discarding burn-in). Parallel chains will be run from different starting values to confirm that the model converges to similar estimates. Other convergence diagnostics will include trace and cumulative distribution plots. Finally, posterior predictive survival will be compared to that which was observed visually and via Bayesian p-values, as was done in previous chapters.

4.1.4 Discussion

In this section, we have proposed a Bayesian hierarchical gamma frailty model with subject-specific parameters for time-to-event data when an individual's risk is the result of an observed number of independent, competing risk processes. The proposed model reflects the biological risk mechanism in this setting by varying the frailty distribution for each exposure process according to individual- and exposure process- specific characteristics. In

this framework, risk-related covariates have a stochastic relationship with frailty, which allows for unobserved heterogeneity as well.

This situation can also be viewed as one with clustered time-to-event data where the event is observed on the cluster level and it is unknown which unit experienced the event. In this context, each exposure process is a unit that can potentially experience an event and exposure processes are clustered within an individual, the level at which events are observed. While existing frailty methods for clustered survival data do not use unit-level covariates to model frailty, the proposed model uses information available in both cluster- and unit-level covariates to model risk.

The proposed model is appropriate for survival data when the number of exposure processes is observed. While the model uses information and exposure processes to model frailty, this approach also has limitations. One limitation of this model is that it assumes the number of exposure processes is measured without error, which may not always be the case in practice. For example, in the MESA data, multiple CT scans were performed for a subset of participants and the number of lesions identified for a participant was not always the same between scans. In our application, we will randomly select a scan to be used in the analysis, but there may be alternative approaches to handling uncertainty in the number of exposure processes while still using exposure process-level data.

Furthermore, the proposed model may insufficiently reflect all sources of risk in some applications. For example, individuals with no lesions detected by CT still have some, albeit small, risk of a CHD event. This limitation must be taken into account when discussing results. In this application, the HR should be interpreted as the individual-level effect on risk of a CHD event for individuals with a lesion detected by CT. If it is important to the scientific question at hand to allow for at-risk individuals with no observed exposure processes in the model, then the method proposed in this chapter would have to be amended to accommodate that. One possible solution would be to first assign all individuals (even those without any exposure processes observed) a single exposure process with frailty dependent on individual-level risk factors. Next, all observed exposure processes, for individuals with at least one observed, would be in addition to this first exposure process.

Finally, it may be desirable to model the number of exposure processes as an observed

Poisson random variable with mean dependent on individual-level covariates. This could be done with a generalized linear regression model of characteristics on the mean parameter and would enable estimation of the effects of some individual-level covariates on the mean number of exposure processes. Data from participants with no exposure processes observed would also be able to contribute to this model.

4.2 Future Work

We conclude with a discussion of some future areas for research.

First, additional research is needed on how to best quantify predictive accuracy for the compound Poisson frailty and compound Poisson mixture frailty models. As mentioned in Section 2.2.3, measures of model fit typically used for Bayesian hierarchical models, DIC and WAIC , do not perform well in latent variable models unless latent variables are integrated out, which is difficult and computationally expensive [16, 37, 46, 52]. Alternative methods based on the posterior predictive distribution, such as the L-measure and predictive concordance, do not accommodate a cured fraction in survival models [20, 27].

Since no criterion appropriate for a latent variable model for survival data with a probability of no risk were available, Bayesian p-values were used throughout this dissertation. In simulations, Bayesian p-values for survival at six-month intervals provided a good indicator of how predicted survival compared to observed survival. In application to HIV prevention trials, however, event times were interval censored so that, consequently, there was no clear observed survival against which to compare predicted survival. In practice, Bayesian p-values are not an adequate measure of model fit for a number of reasons. First, *test quantities*, functions of unknown parameters and data to be compared, are arbitrarily specified. In this dissertation, the test quantity we chose was survival within six month intervals as there is not a universal test quantity for Bayesian survival analysis. If we had chosen a different test quantity, our assessment of model fit may have been different. Next, calculating multiple Bayesian p-values per model does not give a single numeric measure of predictive accuracy; there are as many Bayesian p-values as test quantities specified. Additionally, Bayesian p-values do not account for over-fitting (as is done in DIC, WAIC, and other common criteria). Finally, while values near 0.50 are optimal, there is no consensus

on what Bayesian p-values indicate poor model fit. For all these reasons, Bayesian p-values do not provide an objective measure of predictive accuracy.

A possible alternative for measuring model fit would be to quantify the distance between observed and predicted hazard functions. The Hellinger distance, for example, is a type of f-divergence used to measure the distance between two probability distributions [41]. When comparing predictive accuracy of multiple candidate models, the model with the smallest Hellinger distance provides the best fit to the data. As with Bayesian p-values, however, this measure does not include a penalty for over-fitting the data. Further research is needed to develop a summary measure of model fit for this context.

Second, a limitation of the compound Poisson frailty and compound Poisson mixture frailty models as presented is that the priors on the proportion at risk, $1 - \exp(-\rho)$, and the shape parameter for the distribution of risk associated with each exposure process, η, are skewed. That is, it is not possible to center the prior distribution for either parameter around an expected value while, at the same time, allow for sufficient variability in the prior. The effects of this were seen in simulations; estimates of each parameter were biased, at least in part, due to skewed priors. Developing hierarchical priors for frailty model parameters may be able to address this problem. For example, one could start, as in Chapters 2 and 3, by placing a $Beta(A, B)$ prior on the proportion at risk where $A = PR/(1 - PR) \times B$ so that the expected proportion at risk under the prior is PR. Then, a symmetrical hyperprior can be specified on the expected proportion at risk- for example, $\boldsymbol{\pi}(PR) = Normal(M_{PR}, \Sigma^2_{PR})-$ so that the expected proportion at risk varies but it's prior distribution is still centered at the desired value. Finally, a normal hyperprior can also be placed on B- $\boldsymbol{\pi}(B) = Normal(\mu_B, \Sigma^2_B)$- where M_B is relatively large value so that the Beta prior on the proportion at risk is more symmetrical. Additional exploration of hierarchical priors for these models could improve model fit and make estimates more robust to prior specification.

BIBLIOGRAPHY

[1] Odd O. Aalen. Heterogeneity in survival analysis. *Statistics in Medicine*, 7:1121–1137, 1988.

[2] Odd O. Aalen. Modeling heterogeneity in survival analysis by the compound Poisson distribution. *The Annals of Applied Probability*, 2:951–972, 1992.

[3] Odd O. Aalen and Steinar Tretli. Analyzing incidence of testis cancer by means of a frailty model. *Cancer Causes and Control*, 10:285–292, 1999.

[4] Quarraisha Abdool Karim, Salim S. Abdool Karim, Janet A. Frohlich, Anneke C. Grobler, Cheryl Baxter, Leila E. Mansoor, Ayesha B. M. Kharsany, Sengeziwe Sibeko, Koleka P. Mlisana, Zaheen Omar, Tanuja N. Gengiah, Silvia Maarschalk, Natasha Arulappan, Mukelisiwe Mlotshwa, Lynn Morris, and Douglas Taylor. Effectiveness and safety of tenofovir gel, an antiretroviral microbicide for the prevention of HIV infection in women. *Science*, 329:1168–1174, 2010.

[5] Salim S. Abdool Karim, Barbra A. Richardson, Gita Ramjee, Irving F. Hoffman, Zvavahera M. Chirenje, Taha Taha, Muzala Kapina, Lisa Maslankowski, Anne Coletti, Albert Profy, Thomas R. Moench, Estelle Piwowar-Manning, Benôit Mâsse, Sharon L. Hillier, and Lydia Soto-Torres. Safety and effectiveness of BufferGel and 0.5% PRO2000 gel for the prevention of HIV infection in women. *AIDS*, 25:957–966, 2011.

[6] Arthur S. Agatston, Warren R. Janowitz, Frank J. Hildner, Noel R. Zusmer, Manuel Viamonet, and Robert Detrano. Quantification of coronary artery calcium using ultrafast computed tomography. *Journal of the American College of Cardiology*, 15:827–832, 1990.

[7] Alan Agresti. *Categorical Data Analysis*. John Wiley and Sons, Hoboken, New Jersey, US, 2012.

[8] Jared M. Baeten, Deborah Donnell, Patrick Ndasse, Nelly R. Mugo, James D. Campbell, Jonathan Wangisi, Jordan W. Tappero, Elizabeth A. Bukusi, Craig R. Cohen, Elly Katabira, Allan Ronald, Elioda Tumwesigye, Edwin Were, Kenneth H. Fife, James Kiarie, Carey Farquhar, Grace John-Stewart, Aloysious Kakia, Josephine Ododyo, Akasiima Mucunguzi, Edith Nakku-Joloba, Rogers Tweigye, Kenneth Ngure, Cosmas Apaka, Harrison Tamooh, Fridah Gabona, Andrew Mujugira, Dana Panteleegg, Katherine K. Thomas, Lara Kidoguchi, Meighan. Krows, Jennifer Revall, Susan Morrison, Harald Haugen, Mira Emmanuel-Ogier, Lisa Ondrejcek, Robert W. Coombs, Lisa

Frenkel, Craig Hendrix, Namandjé N. Bumpus, David Bangsberg, Jessica E. Haberer, Wendy S. Stevens, Jairam R. Lingappa, and Connie Celum. Antiretroviral prophylaxis for HIV prevention in heterosexual men and women. *New England Journal of Medicine*, 367:399–410, 2012.

[9] Nikolaus Becker and Werner Rittgen. Some mathematical properties of cumulative damage models regarding their application in cancer epidemiology. *Biometrical Journal*, 32:1–15, 1990.

[10] Joseph Berkson and Robert P. Gage. Survival cure for cancer patients following treatment. *Journal of the American Statistical Association*, 47:501–515, 1952.

[11] Diane E. Bild, David A. Bluemke, G.L. Burke, Robert Detrano, Ana V. Diez Roux, Aaron R. Folsom, Philip Greenland, David R. Jr. Jacobs, Richard Kronmal, Kiang Liu, Jennifer Clark Nelson, Daniel O'Leary, Mohammed F. Saad, Steven Shea, Szklo Moyses, and P. Tracy Russel. Multi-Ethnic Study of Atherosclerosis: Objectives and design. *American Journal of Epidemiology*, 156:871–881, 2002.

[12] Elizabeth R. Brown and Joseph G. Ibrahim. A Bayesian semiparametric joint hierarchical model for longitudinal and survival data . *Biometrics*, 59:221–228, 2003.

[13] Elizabeth R. Brown, Richard A. Kronmal, David A. Bluemke, Alan D. Geurci, J. Jeffrey Carr, Jonathan Goldin, and Robert Detrano. Coronary calcium coverage score: Dextermination, correlates and predictive accuracy in the Multi-Ethnic Study of Atherosclerosis. *Radiology*, 247:669–675, 2008.

[14] Bradley P. Carlin and Thomas A. Louis. *Bayesian Methods for Data Analysis*. Chapman and Hall/CRC, Boca Raton, Florida, 2009.

[15] J. Jeffrey Carr, Jennifer Clark Nelson, Nathan D. Wong, Michael McNitt-Gray, Yadon Arad, David R. Jr. Jacobs, Stephan Sidney, Diane E. Bild, O. Dale Williams, and Robert C. Detrano. Calcified coronary artery plaque measurement with cardiac CT in population-based studies: Standardized protocol of Multi-ethnic study of atherosclerosis (MESA) and Coronary artery risk development in youth (CARDIA) study. *Radiology*, 234:35–43, 2005.

[16] Gilles Celeux, Florence Forbes, Christian P. Robert, and D.M. Titterington. Deviance information criteria for missing data models. *Bayesian Analysis*, 1:651, 2006.

[17] Connie Celum, Anna Wald, James Hughes, Jorge Sanchez, Stewart Reid, Sinead Delany-Moretlwe, Frances Cowan, Martine Casapia, Abner Oritz, Jonathan Fuchs, Susan Buchbinder, Beryl Koblin, Sheryl Zwerski, Scott Rose, Jing Wang, and Lawrence Corey. Effect of aciclovir on HIV-1 acquisition in herpes simplex virus 2 seropositive women and men who have sex with men: a randomised, double-blind, placebo-controlled trial. *Lancet*, 371:2109–19, 2008.

[18] David R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220, 1972.

[19] Chris Elbers and Geert Ridder. True and spurious duration dependence: the identifiability of the proportional hazard model. *The Review of Economic Studies*, 49:403–409, 1982.

[20] Alan Gelfand. Model determination using sampling baed methods. In W. Gilks, S. Richardson, and D. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 145–162. Chapman and Hall/CRC, Boca Raton, Florida, 1996.

[21] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–440, 1990.

[22] Andew Gelman, John B. Carlin, Hal S. Stern, David D. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, London, 2013.

[23] Alan S. Go, Dariush Mozaffarian, Véronique L. Roger, Emelia J. Benjami, Jarret D. Berry, Michael J. Blaha, Shifan Dai, Earl S. Ford, Caroline S. Fox, Sheila Franco, heather J. Fullerton, Cathleen Gillespie, John A. Hailpern, Susan M. an Heit, Virginia J. Howard, Mark D. Huffman, Suzanne E. Judd, Brett M. Kissela, Steven J. Kittner, Daniel T. Lackland, Judith H. Lichtman, Lynda D. Lisabeth, Rachel H. Mackey, David J. Magid, Gregory M. Marcus, Ariane Marelli, David B. Matchar, Carren K. McGuire, Emile R. III Mohler, Claudia S. Moy, Michael E. Mussolino, Robert W. Neumar, Graham Nichol, Dilip K. Pandey, Nina P. Paynter, Matthew J. Reeves, Paul D. Sorlie, Joel Stein, Amytis Towfighi, Tanya N. Turan, Salim S. Virani, Nathan D. Wong, Daniel Woo, and Melanie B. Turner. Heart disease and stroke statistics- 2014 update. *Circulation*, 129:e28–e292, 2014.

[24] Miguel Àngel Hernàn, Babette Brumback, and James M. Robins. Marginal structural models to estimate the causal effect of Zidovudine on the survival of HIV-Ppositive men. *Epidemiology*, 11:561–570, 2000.

[25] Philip Hougaard. Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 71:75–83, 1984.

[26] Philip Hougaard. Frailty models for survival data. *Lifetime Data Analysis*, 1:255–273, 1995.

[27] Joseph G. Ibrahim, Ming-Hui Chen, and Sinha Debajyoti. *Bayesian Survival Analysis*. Springer, New York, 2010.

[28] Erin M. Kahle, James P. Hughes, Jairam R. Lingappa, Grace John-Steward, Connie Celum, Edith Nakku-Joloba, Stella Njuguna, Nelly Mugo, Elizabeth Bukusi, Rachel

Manongi, and Jared M. Baeten. An empiric risk scoring tool for identifying high-risk heterosexual HIV-1 serodiscordant couples for targeted HIV-1 prevention. *Journal of Acquired Immune Deficiency Syndrome*, 62:339–347, 2013.

[29] Mariya Kuk, Thapat Wannarong, Vadim Beletsky, Grace Parraga, Aaron Fenter, and J. David Spence. Volume of carotid artery ulceration as a predictor of cardiovascular events. *Stroke*, 45:1437–1441, 2014.

[30] Dandan Liu, John D. Kalbflesch, and Douglas E. Schaubel. A positive stable frailty model for clustered failure time data with covariate-dependent frailty. *Biometrics*, 67:8–17, 2011.

[31] Ira M. Longini and M. Elizabeth Halloran. A frailty mixture model for estimating vaccine efficacy. *Journal of the Royal Statistical Society, Series C*, 45:165–173, 1996.

[32] Kenneth G. Manton, Eric Stallard, and James W. Vaupel. Methods for comparing the mortality experience of heterogeneous populations. *Demography*, 18:389–410, 1981.

[33] Jeanne M. Marrazzo, Gita. Ramjee, Gonsagrie Nair, Thesla. Palanee, Baningi Mkhize, Clemensia. Nakabiito, Marthinette Taljaard, Jeanna Piper, K. Gomez Feliciano, and Z. Mike Chirenje. Pre-exposure prophylaxis for HIV in women: daily oral tenofovir, oral tenofovis, oral tenofovir/emtricitabine or vaginal tenofovir gel in the VOICE study (MTN 003). In *Conference on Retroviruses and Opportunistic Infections*, 2013.

[34] Mark Mascolini. HIV prevention fails in all three VOICE arms, as daily Truvada PrEP falls. In *Conference on Retroviruses and Opportunistic Infections*, 2013.

[35] Sheena McCormack, Gita Ramjee, Anatoli Kamali, Helen Rees, Angela M. Crook, Mitzy Gafos, Ute Jentsch, Robert Pool, Maureen Chisembele, Saidi Kapiga, Richard Mutemwa, Andrew Vallely, Thesla Palanee, Yuki Sookrajh, Charles J. Lacey, Janet Darbyshire, Heiner Grosskurth, Albert Profy, Andrew Nunn, Richard Hayes, and Jonathan Weber. PRO2000 microbicide gel for prevention of HIV-1 infection (Microbicides Development Programme 301): A phase 3, randomised, double-blind, parallel-group trial. *Lancet*, 376:1329–1337, 2010.

[36] Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42 (2):109–42, 1980.

[37] Russell B. Millar. Comparison of Hierarchical Bayesian Models for Overdispersed Count Data using DIC and Bayes' Factors. *Biometrics*, 65:962–969, 2009.

[38] Tron A. Moger and Odd O. Aalen. A distribution for multivariate frailty based on the compound Poisson distribution with random scale. *Lifetime Data Analysis*, 11:41–59, 2005.

[39] Tron A. Moger, Odd O. Aalen, Ketil Heimdal, and Hakon K. Gjessing. Analysis of testicular cancer data using a frailty model with familial dependence. *Statistics in Medicine*, 23:617–632, 2004.

[40] Radford M. Neal. Slice sampling. *Annals of statistics*, 31:705–741, 2003.

[41] M.A. Nikulin. Hellinger distance. In M. Hazewinkel, editor, *Encyclopedia of Mathematics*. Springer, Berlin, 2001.

[42] Justin J. O'Hagan, Miguel A. Hernàn, Rochelle P. Walensky, and Marc Lipsitch. Apparent declining efficacy in randomized trials: examples of the Thai RV144 HIV vaccine and South African CAPRISA 004 microbicide trials. *AIDS*, 26:123–126, 2012.

[43] Dionne L. Price and Amita K. Manatunga. Modeling survival data with a cured fraction using frailty models. *Statistics in Medicine*, 20:1515–1527, 2001.

[44] James M. Robins, Miguel Àngel Hernàn, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemology*, 11:550–560, 2000.

[45] Josemar Rodrigues, Vincente Cancho, Mario de Castro, and N. Balakrishnan. A Bayesian destructive weighted Poisson cure rate model and an application to a cutaneous melanoma data. *Statistical Methods in Medical Research*, 21:585–597, 2012.

[46] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika van der Linde. Bayesian methods of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64:583–639, 2002.

[47] Michael C. Thigpen, M. Kebaabetswe Poloko, Lynn A. Paxton, Smith Dawn K., Charles E. Rose, Tebogo M. Segolodi, Faith L. Henderson, Sonal R. Pathak, Fatma A. Soud, Kata L. Chillag, Rodreck Mutanhaurwa, Lovemore Ian Chirwa, Michael Kasonde, Daniel Abebe, Evans Buliva, Roman J. Gvetadze, Sandra Johnson, Thom Sukalac, Vasavi T. Thomas, Clyde Hart, Jeffrey A. Johnson, Kevin Malotte, Craig W. Hendrix, and John T. Brooks. Antiretroviral preexposure prophylaxis for heterosexual HIV transmission in Botswana. *New England Journal of Medicine*, 367:423–34, 2012.

[48] A. Tsodikov. Estimation of a survival based on proportional hazards when cure is a possibility. *Mathematical and Computer Modeling*, 33:1227–1236, 2001.

[49] Lut Van Damme, Amy Corneli, Khatija Ahmed, Kawango Agot, Johan Lombaard, Saidi Kapiga, Mookho Malahleha, Fredrick Owino, Rachel Manongi, Jacob Onyango, Lucky Temu, Modie Constance Monendi, Paul Mak'Oketch, Mankalimeng Makanda, Ilse Reblin, Shumani Elsie Makatu, Lisa Saylor, Haddie Kernan, Stella Kerkendale, Christina Wong, Robert Grant, Angela Kashuba, Kavita Nanda, Justin Mandala, Katrien Fransen, Jennifer Deese, Tania Crucitti, Timothy D. Mastro, and Douglas Taylor.

Preexposure prophylaxis for HIV infection among African women. *New England Journal of Medicine*, 367:411–22, 2012.

[50] James W. Vaupel, Kenneth G. Manton, and Eric Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16:439–454, 1979.

[51] James W. Vaupel and Anatoli. I. Yashin. *The deviant dynamics of death in heterogeneous populations*. International Institute for Applied Systems Analysis, Laxenburg, Austria, 1983.

[52] Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely acceptable information criterion in singular learning theory. *The Journal of Machine Learning Research*, 11:3571–3594, 2010.

[53] Michael S. Wulfsohn and Anastasios A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53:330–339, 1997.

Appendix A

APPENDIX FOR CHAPTER 1

Calculations of the population-averaged hazard ratios under gamma and compound Poisson mixing distributions shown in Figure 1 follow directly from Aalen (1988) and are detailed here. (As a clarifying note, these calculations rely on parameterizations of the Cox proportional-hazards model and frailty distributions that differ from the parameterizations used throughout the rest of this dissertation.)

Let $Z\lambda(t)$ be the hazard function for an individual, where $\lambda(t)$ is the average hazard function and the random variable Z is a frailty, a measure of risk, that shifts the hazard up or down, as appropriate. The distribution of Z is the distribution of heterogeneity in risk, referred to here as a mixing distribution. Z has a mixing distribution with mean equal to one and variance equal to ψ, which is also equal to the squared coefficient of variance. This hazard function for an individual is not observed. Instead, we observe the population-level hazard:

$$\mu(t) = \frac{\lambda(t)}{(1 + \psi\frac{1}{\omega}\Lambda(t))^{\omega}} \tag{A.1}$$

where $\mu(t)$ is the hazard function for a randomly selected individual from the population, $\Lambda(t)$ is the cumulative hazard function at time t, and ω is a parameter that specifies the mixing distribution. The two classes for the mixing distribution of Z considered in this chapter are defined as follows:

Gamma distribution ($\omega = 1$): The gamma distribution is the most common choice for the mixing distribution of Z. Z is continuously distributed over the positive half-line, implying that all individuals have some positive risk of an event.

Compound Poisson distribution ($\omega > 1$): For Z that follows a compound Poisson distribution, Z is either equal to zero with positive probability or continuously distributed

on the positive half-line. This form of heterogeneity reflects that some individuals are not susceptible to the disease, condition, or event in question. The proportion of non-susceptible individuals in a population is an increasing function of ω and ψ:

$$P(Z=0) = \exp\Big(-\frac{\omega}{\psi(\omega-1)}\Big), \qquad \omega > 1. \tag{A.2}$$

The distribution of Z among those who are susceptible also varies with ω and ψ. (Note, we introduce a more intuitive hierarchical parameterization for the compound Poisson model in Chapter 2 which we use throughout this dissertation.)

We assume that the average hazard functions for two populations are $\lambda(t)$ and $r\lambda(t)$ where r is a constant less than one, representing a factor that decreases risk of an event proportionally for individuals in the second population. Also assume that the mixing distribution, specified by ψ and ω, is the same in both populations. The hazard ratio (HR) under the Cox model is a ratio of the population-level hazards:

$$\begin{aligned} \frac{\mu_2(t)}{\mu_1(t)} &= \Big(\frac{r\lambda(t)}{(1+\psi\frac{1}{\omega}r\Lambda(t))^{\omega}}\Big) \Big/ \Big(\frac{\lambda(t)}{(1+\psi\frac{1}{\omega}\Lambda(t))^{\omega}}\Big) \\ &= r\Big(\frac{1+\psi\frac{1}{\omega}\Lambda(t)}{1+\psi\frac{1}{\omega}r\Lambda(t)}\Big)^{\omega} \end{aligned} \tag{A.3}$$

We see in (A.3) that the HR under the Cox model would equal the true HR for an individual, r, if $\psi = 0$, i.e. there was no heterogeneity in risk. Also, when t is relatively small, the population HR is close to r. As time increases, however, the population HR increases, approaching the limit $r^{1-\omega}$. When $\omega > 1$, heterogeneity follows the compound Poisson distribution and the limiting value of the population HR is greater than one, indicating increased risk. When $\omega = 1$, heterogeneity follows a gamma mixing distribution and the HR goes to one, indicating no effect.

Results shown in Figure 1.1 were obtained by estimating the population HR as shown in Equation (A.3) and using the following parameter specifications. Population HRs for both mixing distributions were estimated for three variance values, ψ =0.5, 1, 5. The gamma distribution was specified with $\omega = 1$ and the compound Poisson distribution was specified with $\omega = 5$. The proportion at risk given ψ and ω for the compound Poisson model was

calculated via (A.2). The baseline hazard was fixed at 25 events per 100 person-years, $\lambda(t) = 0.25 \, \forall \, t \in (0, 5)$.

Appendix B

APPENDIX FOR CHAPTER 2

We obtain the conditional posterior distribution for each parameter and latent variable from the joint posterior density, Equation (2.5). The conditional log-density of the posterior distributions for β_p, $p = 1, \ldots, P$, ρ, and η to be used in the Gibb's Sampler are

$$\begin{aligned}
\log p(\beta_p|\cdot) &= \Big(\sum_{i=1}^{n} \big(\delta_i (x_{pi}\beta) - Z_i \exp(x_{pi}\beta)\Lambda(t_i)\,\big)\mathbf{1}_{[Z_i>0]} \Big) + \log \pi(\beta_p | M_{\beta_p}, \Sigma^2_{\beta_p}) \\
\log p(\rho|\cdot) &= \Big(\sum_{i=1}^{n} \big(N_i(\log\rho + \eta\log(\rho\eta)) - Z_i\rho\eta\big)\mathbf{1}_{[N_i>0]} - \rho \Big) + \log \pi(\rho|A, B) \\
\log p(\eta|\cdot) &= \Big(\sum_{i=1}^{n} \big(N_i\eta(\log(\rho\eta) + \log(Z_i) - Z_i) - \log\Gamma(N_i\eta)\big)\mathbf{1}_{[N_i>0]} \Big) + \log \pi(\eta|O_\eta, T_\eta)
\end{aligned}$$

where $(\cdot)$ represents the parameters, latent variables, and data being conditioned on. We use a slice sampling algorithm to sample from these conditional posterior distributions.

The conditional log-posterior density for the parameters in $\lambda(t)$ depend on the form specified for the baseline hazard function. Under a constant baseline hazard, $\lambda(t) = \lambda \,\forall\, t$, the gamma distribution is a conjugate prior. With $\boldsymbol{\pi}(\lambda) = Gamma(O_\lambda, T_\lambda)$, the conditional posterior distribution for λ is proportional to that of a gamma random variable with shape parameter $O_\lambda + \sum_{i=1}^{n} \delta_i$ and rate parameter $T_\lambda + \sum_{i=1}^{n} Z_i e^{\mathbf{x}_i'\boldsymbol{\beta}} t_i$. If a more general form is chosen, e.g. $\lambda(t) = \sum_{j=1}^{J} B_j h(t)$, the conditional-log posterior density for each parameter B_j, $j = 1, \ldots, J$ is

$$\log p(B_j|\cdot) = \big(\sum_{i=1}^{n} \delta_i \log(B_j h(t)) - Z_i e^{\mathbf{x}_i'\boldsymbol{\beta}} \int_0^{t_i} \sum_{j=1}^{J} B_j h(t)\, dt \,\big) + \log \pi(B_j | O_{B_j}, T_{B_j}).$$

The specific sampling procedure for each $N_i^{(m)}, Z_i^{(m)}$ for $i = 1, \ldots, n$ at each m^{th} iteration of the sampling algorithm in $m = 2, \ldots, B$ is as follows: (Sampling notation, i.e. $^{(m)}$, is suppressed for all parameters. The current estimate of the parameters would be used in each step.)

1. Sample $N_i^{(m)}$ using the conditional posterior density of $N_i^{(m)}$ marginalized over Z, given in (2.6).

   (a) If $\delta_i = 0$, draw from the specified Poisson distribution

   $$N_i^{(m)}|\delta_i = 0 \sim \text{Poisson}(\rho\Big(\frac{\rho\eta}{\rho\eta + \exp(\mathbf{x}_i'\boldsymbol{\beta})\Lambda(t_i)}\Big)^{\eta})$$

   If $N_i^{(m)} = 0$, set $Z_i^{(m)} = 0$ and exit algorithm.
   If $N_i^{(m)} > 0$, move to step 2 of algorithm.

   (b) If $\delta_i = 1$, use a Metropolis-Hastings algorithm to sample $N_i^{(m)}$.
   Sample candidate value $\check{N}_i$ from a discrete Uniform distribution conditional on current value, $N_i^{(m-1)}$.

   $$p(\check{N}_i|\delta_i = 1, N_i^{(m-1)} = 1) = \begin{cases} N_i^{(m-1)} & \text{w.p. } 1/2 \\ N_i^{(m-1)} + 1 & \text{w.p. } 1/2 \end{cases} \quad \text{(B.1)}$$

   $$p(\check{N}_i|\delta_i = 1, N^{(m-1)} > 1) = \begin{cases} N_i^{(m-1)} - 1 & \text{w.p. } 1/3 \\ N_i^{(m-1)} & \text{w.p. } 1/3 \\ N_i^{(m-1)} + 1 & \text{w.p. } 1/3 \end{cases} \quad \text{(B.2)}$$

   Take

   $$N_i^{(m)} = \begin{cases} \check{N}_i & \text{w.p. } \varrho(N_i^{(m-1)}, \check{N}_i) \\ N_i^{(m-1)} & \text{w.p. } 1 - \varrho(N_i^{(m-1)}, \check{N}_i) \end{cases}$$

   The acceptance probability $\varrho(\cdot, \cdot)$ is

   $$\varrho(N_i^{(m-1)}, \check{N}_i) = \min\left(\frac{f(\check{N}_i|\cdot)}{f(N_i^{(m-1)}|\cdot)} \times \frac{p(N_i^{(m-1)}|\check{N}_i)}{p(\check{N}_i|N_i^{(m-1)})}, 1\right)$$

   where the conditional posterior density $f(N_i|\cdot)$ is given in (2.6) and the sampling probability of a candidate $\check{N}_i$ given $N_i^{(m-1)}$, $p(\cdot|\cdot)$, is given in (B.1) and (B.2).

2. For $N_i^{(m)} > 0$ found in Step 1, sample $Z_i^{(m)}$ using conditional posterior density of Z_i:

   $$\begin{aligned} f(Z_i|\cdot) &\propto \frac{\rho\eta^{N_i\eta} Z_i^{N_i\eta-1}}{\Gamma(N_i\eta)} \exp(-Z_i\rho\eta)\big(\lambda(t)Z_i\exp(\mathbf{x}_i'\boldsymbol{\beta})\big)^{\delta_i} \exp\big(Z_i\exp(\mathbf{x}_i'\boldsymbol{\beta})\Lambda(t_i)\big) \\ &\propto Z_i^{(N_i\eta-1+\delta_i)} \exp\big(-Z_i(\rho\eta + \exp(\mathbf{x}_i'\boldsymbol{\beta})\Lambda(t_i)\big) \end{aligned} \quad \text{(B.3)}$$

We see in (B.3) that the conditional posterior distribution of Z_i is a Gamma distribution:

$$Z_i^{(m)}|\cdot \sim \text{Gamma}\Big(\text{ shape} = (N_i^{(m)}\eta + \delta_i),\ \text{rate} = \big(\rho\eta + \exp(\mathbf{x}_i'\boldsymbol{\beta})\Lambda(t_i)\big)\Big)$$

Appendix C

APPENDIX FOR CHAPTER 3

C.1 Gibb's Sampler

C.1.1 Posterior Samples for Model Parameters

The conditional posterior density for each model parameter and latent variable is obtained from the full joint posterior density in Equation (3.4).

The conditional log-posterior density for each β_p, the log hazard ratio associated with covariate x_p, $p = 1, \ldots, P$, is:

$$\log p(\beta_p|\cdot) = \Big(\sum_{i=1}^{n} \big(\delta_i(x_{pi}\beta_p) - Z_i\, e^{x_{pi}\beta_p}\Lambda(t_i)\,\big)\mathbf{1}_{[N_i>0]}\Big) + \log \pi(\beta_p|M_{\beta_p}, \Sigma^2_{\beta_p})$$

where $(\cdot)$ represents the parameters and data being conditioned on and $(M_{\beta_p}, \Sigma^2_{\beta_p})$ are the parameters for the prior distribution on β_p, $p = 1, \ldots, P$.

The conditional posterior distribution for the parameter(s) in $\lambda(t)$ depend on the form chosen for the baseline hazard function. Under a piecewise constant baseline hazard, $\lambda_0(t) = \lambda_{0s}$, $t_s \le t < t_{s+1}$, $s = 1, \ldots, S$, the gamma distribution is a conjugate prior for each λ_{0s}: $\boldsymbol{\pi}(\lambda) = Gamma(O_\lambda, T_\lambda)$. Then, the posterior density for each λ_{0s} is proportional to that of a gamma random variable with shape parameter $O_\lambda + \sum_{i=1}^{n}(\delta_i \mathbf{1}_{t_s \le t_i < t_{s+1}})$ and rate parameter $T_\lambda + \sum_{i=1}^{n} Z_i e^{\mathbf{x}_i'\boldsymbol{\beta}} \sum_i^n \big((t_i - t_s)\mathbf{1}_{t_s \le t_i < t_{s+1}} + (t_{s+1} - t_s)\mathbf{1}_{t_i \ge t_{s+1}}\big)$.

The conditional log-posterior densities for the mean number of exposure processes, ρ_k, and the shape parameter for the gamma-distributed random variable associated with each exposure process, η_k, for each risk class k, $k = 1, \ldots, K$, are:

$$\log p(\rho_k|\cdot) = \sum_{i=1}^{n} \Big(\big(N_i \log\rho_k - \rho_k\big)\mathbf{1}_{[R_i=k]} + \sum_{l=1}^{K} \big(N_i\eta_l(\log\nu) - Z_i\nu\big)\mathbf{1}_{[R_i=l,\, N_i>0]}\Big) + \log\boldsymbol{\pi}(\rho_k|A_k, B_k)$$

$$\log p(\eta_k|\cdot) = \sum_{i=1}^{n} \Big(\sum_{l=1}^{K} \big(N_i\eta_l(\log\nu - \log Z_i) - \log\Gamma(N_i\eta_l) - Z_i\nu\big)\mathbf{1}_{[R_i=l,\, N_i>0]}\Big) + \log\boldsymbol{\pi}(\eta_k|O_{\eta_k}, T_{\eta_k})$$

The value of ν is estimated at each iteration with the current samples of $\boldsymbol{\rho}$ and $\boldsymbol{\eta}$, as well as cumulative link model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, following the definition given in Equation (3.2): $\nu = \sum_{k=1}^{K} \rho_k \eta_k \frac{1}{n} \sum_{i=1}^{n} \big(G^{-1}(\alpha_k - \mathbf{w}_i'\boldsymbol{\gamma}_k) - G^{-1}(\alpha_{k-1} - \mathbf{w}_i'\boldsymbol{\gamma}_{k-1})\big)$. It follows that the conditional log-posterior for each ρ_k and η_k are conditional on all other $\rho_{k'}$ and $\eta_{k'}$, $k' \neq k$. Each ρ_k and η_k are sampled separately. Ordering of parameter values $\rho_1 < \rho_2 < \cdots < \rho_K$ and $\eta_1 < \eta_2 < \cdots < \eta_K$ is imposed in the sampling process.

Each component of the parameter vector $\boldsymbol{\gamma}_k$, which models the effect of covariates $\mathbf{w}_i$ on classification probabilities, has the conditional log-posterior:

$$\begin{aligned}\log p(\gamma_{k,d}|\cdot) \;=\; & \sum_{i=1}^{n} \Big(\sum_{k=1}^{K} \log\big(G^{-1}(\alpha_k - \mathbf{w}_i'\boldsymbol{\gamma}_k) - G^{-1}(\alpha_{k-1} - \mathbf{w}_i'\boldsymbol{\gamma}_{k-1})\big)\mathbf{1}_{[R_i=k]} \\ & + \sum_{l=1}^{K} \big(N_i\eta_l(\log\nu) - Z_i\nu\big)\mathbf{1}_{[R_i=l,\, N_i>0]}\Big) + \log\boldsymbol{\pi}(\gamma_{k,d}|M_{\gamma_{k,d}}, \Sigma^2_{\gamma_{k,d}})\end{aligned}$$

for $d = 1, \ldots, D$ and $k = 1, \ldots, K-1$. The conditional log-posterior for each component of the parameter of intercepts $\boldsymbol{\alpha}$ is:

$$\begin{aligned}\log p(\alpha_{k^\dagger}|\cdot) \;=\; & \sum_{i=1}^{n} \Big(\sum_{k=k^\dagger}^{k^\dagger+1} \log\big(G^{-1}(\alpha_k - \mathbf{w}_i'\boldsymbol{\gamma}_k) - G^{-1}(\alpha_{k-1} - \mathbf{w}_i'\boldsymbol{\gamma}_{k-1})\big)\mathbf{1}_{[R_i=k]} \\ & + \sum_{l=1}^{K} \big(N_i\eta_l(\log\nu) - Z_i\nu\big)\mathbf{1}_{[R_i=l,\, N_i>0]}\Big) + \log\boldsymbol{\pi}(\alpha_l|M_{\alpha_l}, \Sigma^2_{\alpha_l})\end{aligned}$$

for $k^\dagger = 1, \ldots, K-1$, $\alpha_0 = -\infty$, and $\alpha_K = \infty$. Order for components of $\boldsymbol{\alpha}$ is imposed during the sampling process. Again, $\nu = \sum_{k=1}^{K} \rho_k \eta_k \frac{1}{n} \sum_{i=1}^{n} \big(G^{-1}(\alpha_k - \mathbf{w}_i'\boldsymbol{\gamma}_k) - G^{-1}(\alpha_{k-1} - \mathbf{w}_i'\boldsymbol{\gamma}_{k-1})\big)$.

C.1.2 Sampling Latent Variables

Latent variables, R_i, N_i, and Z_i, $i = 1, \ldots, n$ are also updated at each iteration of the Gibbs sampler. We first derive the conditional densities for each latent variable before detailing the sampling procedure.

For R_i, the latent risk class for participant i, the conditional density is:

$$f(R_i|\cdot) \propto \prod_{k=1}^{K} \Big(\big(G^{-1}(\alpha_k - \mathbf{w}_i'\boldsymbol{\gamma}_k) - G^{-1}(\alpha_{k-1} - \mathbf{w}_i'\boldsymbol{\gamma}_{k-1})\big) \\ e^{-\rho_k} \big(\frac{\rho_k^{N_i}}{\Gamma(N_i+1)} \frac{\nu^{N_i\eta_k}}{\Gamma(N_i\eta_k)} Z_i^{N_i\eta_k} \big)^{\mathbf{1}_{[N_i>0]}} \Big)^{\mathbf{1}_{[R_i=k]}}.$$

We sample the class for each participant using a Metropolis-Hastings type algorithm.

Next, in order to sample N_i, it is necessary to remove Z_i from the conditional density of N_i, as an individual's frailty is dependent on her number of exposure processes, per the hierarchical definition of the compound Poisson distribution. So, we sample N_i from its distribution conditional on observed data and all other parameters but marginalized over Z_i:

$$\begin{aligned} &f(N_i|R_i = k, t_i, \delta_i, \mathbf{x}_i, \boldsymbol{\beta}, \lambda(t), \boldsymbol{\eta}, \boldsymbol{\rho}) \\ &\quad \propto \big(\frac{\rho_k^{N_i}}{\Gamma(N_i+1)} \big) \int_0^\infty \frac{\nu^{N_i\eta_k}}{\Gamma(N_i\eta_k)} Z_i^{N_i\eta_k - 1} e^{-Z_i\nu} \left(\lambda(t) Z_i e^{\mathbf{x}_i'\boldsymbol{\beta}}\right)^{\delta_i} e^{-Z_i e^{\mathbf{x}_i'\boldsymbol{\beta}} \Lambda(t_i)} dZ_i \\ &\quad \propto \frac{\rho_k^{N_i}}{\Gamma(N_i+1)} \Big(\frac{\nu}{\nu + e^{\mathbf{x}_i'\boldsymbol{\beta}}\Lambda(t_i)} \Big)^{N_i\eta_k} \Big(\frac{N_i\eta_k}{\nu + e^{\mathbf{x}_i'\boldsymbol{\beta}}\Lambda(t_i)} \Big)^{\delta_i} \end{aligned} \tag{C.1}$$

The approach for sampling N_i from Equation (C.1) depends on whether seroconversion was observed. When $\delta_i = 0$, i.e. seroconversion was not observed for participant i, the conditional density of N_i marginalized over Z_i is proportional to that of a Poisson distribution with mean $\rho\big(\nu/(\nu + \exp(\mathbf{x}_i'\boldsymbol{\beta})\Lambda(t_i))\big)^{\eta_k}$. We draw from this Poisson to obtain N_i. When $\delta_i = 1$, i.e. seroconversion was observed, we use a Metropolis-Hastings type algorithm to sample N_i from Equation (C.1), with the constraint that N_i must be positive. (Since seroconversion occurred, the participant must have had at least one exposure process.)

After sampling the number of exposure processes for an individual, we sample frailty. If the sampled number of exposure processes for participant i, N_i, is nonzero, the conditional density of Z_i is proportional to the density of a gamma distribution with shape $= N_i\eta_k + \delta_i$ and rate $= \nu + e^{\mathbf{x}_i'\boldsymbol{\beta}}\Lambda(t_i)$, as shown below:

$$\begin{aligned} f(Z_i|\cdot) &\propto \frac{\nu^{N_i\eta_k} Z_i^{N_i\eta_k - 1}}{\Gamma(N_i\eta_k)} e^{-Z_i\nu} \big(\lambda(t) Z_i e^{\mathbf{x}_i\boldsymbol{\beta}}\big)^{\delta_i} e^{Z_i e^{\mathbf{x}_i\boldsymbol{\beta}}\Lambda(t_i)} \\ &\propto Z_i^{(N_i\eta_k - 1 + \delta_i)} e^{-Z_i(\nu + e^{\mathbf{x}_i'\boldsymbol{\beta}}\Lambda(t_i))} \end{aligned} \tag{C.2}$$

Hence, to obtain Z_i conditional on N_i, observed data, and all parameters, we simply draw from this gamma distribution. If the sampled number of exposure processes is zero, then the frailty is necessarily zero.

Thus, we have the following sampling procedure for each $(R_i^{(m)}, N_i^{(m)}, Z_i^{(m)})$, $i = 1, \ldots, n$, at the m^{th} iteration of the sampling algorithm in $m = 2, \ldots, B$. Here, iteration number notation, i.e. $^{(m)}$ is suppressed for all parameters. The current estimates of all parameters are used for each step, including ν, which is defined based on current parameter estimates following the relationship defined in Equation (3.2).

1. To sample latent risk class $R_i^{(m)}$, first sample class $\check{R}_i \in k = 1, \ldots, K$ from a discrete Uniform distribution conditional on the current class, $R_i^{(m-1)}$.

$$p(\check{R}_i |\, R_i^{(m-1)} = 1) = \begin{cases} R_i^{(m-1)} & \text{w.p. } 1/2 \\ R_i^{(m-1)} + 1 & \text{w.p. } 1/2 \end{cases} \tag{C.3}$$

$$p(\check{R}_i |\, 1 < R_i^{(m-1)} < K) = \begin{cases} R_i^{(m-1)} - 1 & \text{w.p. } 1/3 \\ R_i^{(m-1)} & \text{w.p. } 1/3 \\ R_i^{(m-1)} + 1 & \text{w.p. } 1/3 \end{cases} \tag{C.4}$$

$$p(\check{R}_i |\, R_i^{(m-1)} = K) = \begin{cases} R_i^{(m-1)} - 1 & \text{w.p. } 1/2 \\ R_i^{(m-1)} & \text{w.p. } 1/2 \end{cases} \tag{C.5}$$

   Then, take

$$R_i^{(m)} = \begin{cases} \check{R}_i & \text{w.p. } \varrho(R_i^{(m-1)}, \check{R}_i) \\ R_i^{(m-1)} & \text{w.p. } 1 - \varrho(R_i^{(m-1)}, \check{R}_i) \end{cases}$$

   The acceptance probability $\varrho(\cdot, \cdot)$ is

$$\varrho(R_i^{(m-1)}, \check{R}_i) = \min\left(\frac{f(\check{R}_i|\cdot)}{f(R_i^{(m-1)}|\cdot)} \times \frac{p(R_i^{(m-1)}|\check{R}_i)}{p(\check{R}_i|R_i^{(m-1)})}, 1 \right)$$

   where the conditional density of $f(R_i|\cdot)$ is given in Equation (C.1) and the conditional sampling probability $p(\cdot|\cdot)$ is given in Equation (C.3), Equation (C.4), and Equation (C.5).

2. Sample $N_i^{(m)}$ using the conditional density of $N_i^{(m)}$ marginalized over Z, given in Equation (C.1).

   (a) If $\delta_i = 0$, draw from the specified Poisson distribution

   $$N_i^{(m)}|\delta_i = 0 \sim \text{Poisson}\left(\rho_k\Big(\frac{\nu}{\nu + e^{\mathbf{x}_i'\boldsymbol{\beta}}\Lambda(t_i)}\Big)^{\eta_k}\right)$$

   If $N_i^{(m)} = 0$, set $Z_i^{(m)} = 0$ and exit algorithm.
   If $N_i^{(m)} > 0$, move to step 2 of algorithm.

   (b) If $\delta_i = 1$, use a Metropolis-Hastings algorithm to sample $N_i^{(m)}$.
   Sample candidate value $\check{N}_i$ from a discrete Uniform distribution conditional on current value, $N_i^{(m-1)}$.

   $$p(\check{N}_i|\delta_i = 1, N_i^{(m-1)} = 1) = \begin{cases} N_i^{(m-1)} & \text{w.p. } 1/2 \\ N_i^{(m-1)} + 1 & \text{w.p. } 1/2 \end{cases} \tag{C.6}$$

   $$p(\check{N}_i|\delta_i = 1, N^{(m-1)} > 1) = \begin{cases} N_i^{(m-1)} - 1 & \text{w.p. } 1/3 \\ N_i^{(m-1)} & \text{w.p. } 1/3 \\ N_i^{(m-1)} + 1 & \text{w.p. } 1/3 \end{cases} \tag{C.7}$$

   Take

   $$N_i^{(m)} = \begin{cases} \check{N}_i & \text{w.p. } \varrho(N_i^{(m-1)}, \check{N}_i) \\ N_i^{(m-1)} & \text{w.p. } 1 - \varrho(N_i^{(m-1)}, \check{N}_i) \end{cases}$$

   The acceptance probability $\varrho(\cdot, \cdot)$ is

   $$\varrho(N_i^{(m-1)}, \check{N}_i) = \min\left(\frac{f(\check{N}_i|\cdot)}{f(N_i^{(m-1)}|\cdot)} \times \frac{p(N_i^{(m-1)}|\check{N}_i)}{p(\check{N}_i|N_i^{(m-1)})}, 1\right)$$

   where the conditional density $f(N_i|\cdot)$ is given in Equation (C.1) and the conditional sampling probability $p(\cdot|\cdot)$ is given in Equation (C.6) and Equation (C.7).

3. For $N_i^{(m)} > 0$ found in Step 2, sample $Z_i^{(m)}$ from the following Gamma distribution (as demonstrated in Equation (C.2)):

   $$Z_i^{(m)} \sim \text{Gamma}\Big(\text{shape} = (N_i^{(m)}\eta_k + \delta_i), \text{rate} = \big(\nu + e^{\mathbf{x}_i'\boldsymbol{\beta}}\Lambda(t_i)\big)\Big).$$

C.2 Data Generation for Simulations

Survival times were generated for n=3,000 participants under the latent class compound Poisson mixture frailty model outlined in Chapter 3. Observations were divided evenly between treatment and control arms:

$$x_i = \begin{cases} 0 & i = 1, \ldots, 1,500 \\ 1 & i = 1,501, \ldots, 3,000 \end{cases}$$

Covariate values for the proportional odds model were defined as follows. First, covariate A had three separate levels, divided equally among participants in each arm, i.e. 500 at each level in each arm. This covariate translates into two indicator variables, w_1 and w_2:

$$w_{i1} = \begin{cases} 0 & i = 501, \ldots, 1,500;\ 2,001, \ldots, 3,000 \\ 1 & i = 1, \ldots, 500;\ 1,501, \ldots, 2,000 \end{cases}$$

$$w_{i2} = \begin{cases} 0 & i = 1, \ldots, 500;\ 1,001, \ldots, 2,000;\ 2501, \ldots, 3,000 \\ 1 & i = 501, \ldots, 1,000;\ 2,001, \ldots, 2,500 \end{cases}$$

Next, covariate B was defined as a binary variable and positively correlated with covariate A. Among those with $w_1 = 1$, 100 participants in each arm had $w_3 = 1$, among those with $w_2 = 1$, 80 participants in each arm had $w_3 = 1$, and among those with $w_1 = w_2 = 0$, 60 participants in each arm had $w_3 = 1$:

$$w_{i3} = \begin{cases} 0 & i = 101, \ldots, 500;\ 581, \ldots, 1,000;\ 1,061, \ldots, 1,500; \\ & \quad 1,601, \ldots, 2,000;\ 2,081, \ldots, 2,500;\ 2,561, \ldots, 3,000 \\ 1 & i = 1, \ldots, 100;\ 501, \ldots, 580;\ 1,001, \ldots, 1,060; \\ & \quad 1,501, \ldots, 1,600;\ 2,001, \ldots, 2,080;\ 2,501, \ldots, 2,560 \end{cases}$$

Finally covariate C was also defined as a binary variable and positive correlated with covariate A. Here, w_4 is a bernoulli random variable with probability dependent on covariate A:

$$w_{i4} \sim \begin{cases} Bernoulli(0.25) & w_{i1} = 1,\ w_{i2} = 0 \\ Bernoulli(0.5) & w_{i1} = 0,\ w_{i2} = 1 \\ Bernoulli(0.75) & w_{i1} = 0,\ w_{i2} = 0 \end{cases}$$

The same sampled values for w_4 were used for each simulation. (Values were drawn after calling the function `set.seed(1)` in `R`.) We chose to correlate risk-related covariates in order to reflect data one would encounter in actual applications. For example, in HIV prevention we expect age to be associated with risk factors like STIs and marital status.

The multinomial distribution of classification probabilities for each participant was defined by covariates $\mathbf{w}_i$, coefficient vector $\boldsymbol{\gamma}$, and intercept vector $\boldsymbol{\alpha}$ using the proportional odds model:

$$\begin{aligned} P(R_i = 1|\mathbf{w}_i) &= \text{expit}\big(\alpha_1 - \mathbf{w}_i'\boldsymbol{\gamma}\big) \\ P(R_i = 2|\mathbf{w}_i) &= \text{expit}\big(\alpha_2 - \mathbf{w}_i'\boldsymbol{\gamma}\big) - \text{expit}\big(\alpha_1 - \mathbf{w}_i'\boldsymbol{\gamma}\big) \\ P(R_i = 3|\mathbf{w}_i) &= 1 - \text{expit}\big(\alpha_2 - \mathbf{w}_i'\boldsymbol{\gamma}\big) \end{aligned}$$

where expit is the inverse logit function, and $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are defined as follows:

$$\begin{aligned} \boldsymbol{\alpha} &= [-0.5,\ 1] \\ \exp(\boldsymbol{\gamma}) &= [2,\ 1.5,\ 3,\ 0.5]. \end{aligned}$$

Classes were generated for each simulation replication based on these classification probabilities. Frailties were then generated conditional on class, following the compound Poisson mixture model described in Section 3.2.2 and the parameter values

$$\begin{aligned} \boldsymbol{\rho} &= [-\log(0.95),\ -\log(0.75),\ -\log(0.2)] \\ \boldsymbol{\eta} &= [2,\ 3,\ 4] \end{aligned} \tag{C.8}$$

where $\boldsymbol{\rho}$ corresponds to 5%, 25%, and 80% at risk for the low, medium, and high risk classes, respectively. Finally, event times are generated for each participant conditional on sampled frailty, a constant baseline hazard of 5 events per 100 person-years, $\lambda = 0.05$, and a hazard ratio associated with the intervention of 0.5.

Appendix D

APPENDIX FOR CHAPTER 4

The conditional posterior density for each model parameter is obtained from the full joint posterior density in Equation (4.6).

The conditional log-posterior density for each $\boldsymbol{\beta}$, the vector of log relative-risks associated with covariate vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, is:

$$\log p(\boldsymbol{\beta}|\cdot) = \Big(\sum_{i=1}^{n} \big(\delta_i(\mathbf{x}_i'\boldsymbol{\beta}) - Z_i \, e^{\mathbf{x}_i'\boldsymbol{\beta}} \Lambda(t_i) \big) \Big) + \log \pi(\boldsymbol{\beta}|\mathbf{M}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}^{\mathbf{2}}{}_{\boldsymbol{\beta}})$$

where $(\cdot)$ represents the parameters and data being conditioned on and $(\mathbf{M}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}^{\mathbf{2}}{}_{\boldsymbol{\beta}})$ are the mean vector and covariance matrix for the prior distribution on $\boldsymbol{\beta}$, assuming a multivariate normal distribution. (Other prior distributions are certainly possible.)

The conditional posterior distribution for the parameter(s) in $\lambda(t)$ depend on the form chosen for the baseline hazard function. Under a constant baseline hazard, $\lambda(t) = \lambda \, \forall t$, the gamma distribution is a conjugate prior: $\boldsymbol{\pi}(\lambda) = Gamma(O_\lambda, T_\lambda)$. Then, the posterior density for λ is proportional to that of a gamma random variable with shape parameter $O_\lambda + \sum_{i=1}^{n} \delta_i + 1$ and rate parameter $T_\lambda + \sum_{i=1}^{n} Z_i e^{\mathbf{x}_i'\boldsymbol{\beta}} t_i$. This can be extended to a piecewise constant baseline hazard function as well.

Assuming a multivariate normal prior for $\boldsymbol{\xi}$, the parameter vector relating unit- and cluster-level covariates to the shape parameter influencing magnitude of risk for each unit, the conditional log-posterior density for $\boldsymbol{\xi}$ is:

$$\log p(\boldsymbol{\xi}|\cdot) = \sum_{i=1}^{n} \Big(\sum_{m=1}^{N_i} g^{-1}(\mathbf{v}_{im}'\boldsymbol{\xi}) \log\nu - \log\Gamma\big(\sum_{m=1}^{N_i} g^{-1}(\mathbf{v}_{im}'\boldsymbol{\xi})\big) + \sum_{m=1}^{N_i} g^{-1}(\mathbf{v}_{im}'\boldsymbol{\xi}) \log(Z_i) - Z_i\nu \Big) + \log \pi(\boldsymbol{\xi}|\mathbf{M}_{\boldsymbol{\xi}}, \boldsymbol{\Sigma}^{\mathbf{2}}{}_{\boldsymbol{\xi}})$$

where $\nu = \frac{1}{n} \sum_{i=1}^{n} \sum_{m=1}^{N_i} g^{-1}(\mathbf{v}_{im}\boldsymbol{\xi})$ and $(\mathbf{M}_{\boldsymbol{\xi}}, \boldsymbol{\Sigma}^{\mathbf{2}}{}_{\boldsymbol{\xi}})$ are the mean vector and covariance matrix specified for the prior distribution.

It is also necessary to sample partially informed frailty, Z_i, for each cluster $i = 1, \ldots, n$ at each step of the Gibbs sampler. The conditional density of Z_i is proportional to the density of a gamma random variable with shape $\delta_i + \sum_{m=1}^{N_i} g^{-1}(\mathbf{v}'_{im}\boldsymbol{\xi})$ and rate $\nu + \exp(\mathbf{x}'_i\boldsymbol{\beta})\Lambda(t_i)$, as shown below:

$$\begin{aligned} f(Z_i|\cdot) &\propto \frac{\nu^{\sum_{m=1}^{N_i} g^{-1}(\mathbf{v}'_{im}\boldsymbol{\xi})}}{\Gamma(\sum_{m=1}^{N_i} g^{-1}(\mathbf{v}'_{im}\boldsymbol{\xi}))} Z_i^{\sum_{m=1}^{N_i} g^{-1}(\mathbf{v}'_{im}\boldsymbol{\xi})-1} e^{-Z_i\nu} Z_i^{\delta_i} e^{-Z_i e^{\mathbf{x}'_i\boldsymbol{\beta}}\Lambda(t_i)} \\ &\propto Z_i^{(\sum_{m=1}^{N_i} g^{-1}(\mathbf{v}'_{im}\boldsymbol{\xi})-1+\delta_i)} e^{-Z_i(\nu+e^{\mathbf{x}'_i\boldsymbol{\beta}}\Lambda(t_i))} \end{aligned}$$

Hence, to obtain Z_i conditional on model parameters and observed data, we simply draw from this gamma distribution.