Statistical Methods for Assessing Complex Multi-Exposure Data in HIV and Genetic Epidemiology

Statistical Methods for Assessing Complex Multi-Exposure Data in HIV and Genetic Epidemiology

A dissertation presented

by

Katharine Fischer Berry Correia

to

The Department of Biostatistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biostatistics

Harvard University

Cambridge, Massachusetts

April 2018

Dissertation Advisor: Paige Williams Katharine Fischer Berry Correia

Statistical Methods for Assessing Complex Multi-Exposure Data in HIV and Genetic Epidemiology

Abstract

Complex, multi-exposure problems arise in many forms. In this dissertation, we delve into three disparate forms of complex, multi-exposure questions, from the safety of combination antiretroviral (ARV) regimens to the complexities arising from repeated measures data to statistical genetics.

In Chapter 1, we evaluate a hierarchical model that groups ARVs by drug class, while still providing individual ARV effect estimates, to screen for the safety of ARV exposures during pregnancy. In simulations, we compare the statistical operating characteristics of the hierarchical approach to the standard approaches of separate regression models for each ARV and a full, fixed effect model. We illustrate the characteristics of the hierarchical approach in an application evaluating risk of preterm delivery using a study including over 2,000 pregnancies representing over 100 antiretroviral combinations, each involving up to three drug classes.

Chapter 2 explores estimation of the relative excess risk due to interaction (RERI) in clustered data settings. The RERI is a measure of additive interaction for binary outcomes that can be calculated from multiplicative regression models. We evaluate the RERI for the setting of clustered data using both population-averaged and cluster-conditional models. In simulation studies, we find that estimation and inference for the RERI using population-averaged models is straightforward. However, frequentist implementations of cluster-conditional models including

random intercepts often fail to converge or produce degenerate variance estimates. We develop a Bayesian implementation of log binomial random intercept models, which represents an attractive alternative for estimating the RERI in cluster-conditional models. We apply the methods to an observational study of adverse birth outcomes in mothers with HIV infection, in which mothers are clustered within clinical research sites.

In Chapter 3, we introduce a computationally efficient algorithm for permutation testing between a single rare genetic variant and affection status which also allows for adjustment of covariates. To demonstrate the feasibility of the algorithm, we apply the method to a study of chronic obstructive pulmonary disease. In simulations, we show that the permutation test maintains a Type I error rate closer to the nominal level than the asymptotic and saddlepoint approximation tests for rare variants.

Contents

List of Figures

List of Tables

I dedicate this dissertation to my children, Felipe and Olivia, who nourished my spirit during these years with their infectious laughter, inquisitive minds, and compassionate nature.

Acknowledgements

First and foremost, I would like to thank my advisor Paige Williams for her persistent encouragement and guidance over the last five years. Paige, I am grateful for the always prompt and always thorough feedback you gave me on all the papers and presentations I have composed over these years; it was not only beneficial in strengthening the work itself, but invaluable to me personally as your attention and interest in our work instilled confidence in me as a researcher. I am thankful for the support you've afforded me as a doctoral student raising one child – then two! – an hours-plus commute outside the city. You have been a wonderful model of a successful mother, teacher, mentor, researcher, collaborator, and friend.

I also would like to thank Christoph Lange for his faith in my ability to contribute to the field of statistical genetics, with which I had no prior experience before our work together. I much appreciated your always-positive outlook as I vacillated between despair (my simulations aren't working!) and joy (my simulations are working!). Many thanks to Julian Hecker for his guidance and support as I stumbled on basic statistical genetics concepts; I would still be stuck in despair (my simulations aren't working!) had it not been for your instrumental assistance and insights.

Thank you to my committee members Brent Coull and George Seage for your constructive feedback and encouragement that led me to explore new paths of additive interaction (George) and Bayesian approaches (Brent). Concentration in these areas not only strengthened this dissertation, but also broadened my capabilities as a public health researcher.

Lastly, I owe a world of thanks to my original and strongest supporters, my family. Thank you, Mom, for listening to all the ins and outs of my days, and for reminding me that (as Emily Dickenson wrote, since poets say things best)

> "Hope" is the thing with feathers
> That perches in the soul –
> And sings the tune without the words
> And never stops – at all –

Thank you, Dad, for your enduring (and endearing) cheers throughout the years – those sentiments helped propel me through these last five – and for exemplifying the importance of exercise and quietude in cultivating a healthy mind; my husband, Junior, for your unwavering faith in my ability to do anything and everything, and for your renewed support as I struggled to do everything and anything; my brother, Nate, for modeling courage and compassion, and for serving as a humbling reminder of where the real important work lies (hint: it's yours, not mine); my soul sisters Jenn, Sarah, Emily and Erica, for being the most fabulous dessert-loving, Catan-playing, always-there-when-you-need-her crew a woman could ask for; and last, but certainly not least, the next generation – Buddy, Livy, and Emmy Bear – for making our homes full of curiosity, love, and laughter.

Introduction

Complex, multi-exposure problems arise in many forms. In this dissertation, we delve into three disparate forms of complex, multi-exposure questions, from the safety of combination antiretroviral regimens to the complexities arising from repeated measures data to statistical genetics. Each of these statistical challenges is driven by a pressing, underlying clinical question; and ultimately with this dissertation, I aim to provide statistical approaches that clinical researchers can implement in their research for sound statistical practice.

In Chapter 1, we investigate a hierarchical modeling approach for assessing the safety of antiretroviral drug regimens taken during pregnancy by women with HIV. Combination antiretroviral regimens have achieved tremendous success in reducing perinatal HIV transmission, and have become standard of care in pregnant women with HIV. However, the large variety of combination antiretroviral regimens utilized in practice raises the question of whether some of these highly potent drugs pose other risks to the pregnancy or infant. While pregnant women with HIV are almost always exposed to multiple antiretrovirals concurrently, standard safety screening strategies typically consider each individual antiretroviral separately, which fails to account for potential confounding due to simultaneous exposure to other antiretrovirals. We evaluate a hierarchical modeling approach which groups antiretrovirals by drug class, while still providing individual antiretroviral drug effect estimates. We illustrate the characteristics of the hierarchical approach in an application evaluating risk of preterm birth using a study including over 2,000 pregnancies representing over 100 antiretroviral combinations, each involving up to three drug classes.

In addition to screening for adverse individual antiretroviral effects, it is important to consider possible drug-drug and drug-covariate interactions. In particular, for binary outcomes, the risk difference scale and additive interaction effects are often of greater clinical relevance than the ratio scale and multiplicative interaction effects. Yet, the models typically used for binary outcomes implicitly measure interaction on the multiplicative scale. One measure to assess additive interaction from multiplicative models is the Relative Excess Risk due to Interaction (RERI). Extending the hierarchical model assessed in Chapter 1 to screen for additive interactions induces a distribution on the RERIs for each drug-drug or drug-covariate interaction. Furthermore, for common, binary outcomes, it is important to estimate the RERI using relative risks, not odds ratios. Log binomial regression can be unstable, and a hierarchical log binomial regression model with random slopes for each drug and each interaction proved difficult to implement in a frequentist setting.

In Chapter 2, we consider estimating the RERI in more general clustered data settings. The RERI measure has been applied in many contexts, but one limitation of previous approaches is that clustering in data has rarely been considered. We evaluate the RERI metric for the setting of clustered data using both population-averaged and cluster-conditional models. In simulation studies, we find that estimation and inference for the RERI using population-averaged models is straightforward. However, frequentist implementations of cluster-conditional models including random intercepts often fail to converge or produce degenerate variance estimates. We develop a Bayesian implementation of log binomial random intercept models, which represents an attractive alternative for estimating the RERI in cluster-conditional models. We apply the methods to an observational study of adverse birth outcomes in mothers with HIV infection, in which mothers are clustered within clinical research sites.

In Chapter 3, we turn our focus to a setting where the multitude of exposures explodes – statistical genetics. With the rapid advancement in DNA sequencing technologies over the last decade, cost-effective identification of rare and very rare single-nucleotide polymorphisms (SNPs) has become possible. Yet, the standard statistical methods used to test these rare variants rely on asymptotic, large sample theory, which likely does not hold when the minor allele count is so low. However, the computational burden of permutation testing in a logistic regression setting can be prohibitive. We develop a computationally efficient algorithm for permutation testing of individual rare genetic variants that allows for adjustment of covariates. To demonstrate the feasibility of the algorithm, we apply the method to a study of chronic obstructive pulmonary disease. In simulations, we show that the permutation test maintains a Type I error rate closer to the nominal level than the asymptotic and saddlepoint approximation tests.

We conclude this dissertation with a few suggestions for avenues of further research.

1. A hierarchical modeling approach for assessing the safety of exposure to complex antiretroviral drug regimens during pregnancy

The use of combination antiretroviral (ARV) therapy during pregnancy has been a public health success, reducing the risk of perinatal human immunodeficiency virus (HIV) transmission to less than 2% (CDC, 2006; Suksomboon et al., 2007). Despite widespread use of ARVs during pregnancy, there is a dearth of adequate and well-controlled human studies evaluating the safety of ARVs in pregnancy, leading to a need to monitor potential adverse effects that these highly potent drugs may have on the pregnancy or infant (Zash et al., 2016). Given the large number of available and effective ARVs, identification of individual ARVs with increased risks is critical, so that pregnant women can be advised to take ARVs with the safest profile.

The difficulty in assessing the safety of ARVs during pregnancy is due in part to the large number of different drugs available, yielding hundreds of possible combinations of ARV drugs that women can be exposed to during pregnancy. When prior research findings are suggestive or in settings with limited variability in regimens, a comparative effectiveness strategy may be used to compare two regimens against each other (Caniglia et al., 2016). However, such approaches may not be useful for general safety screening across many ARVs or regimens. In most cases, safety screening for a larger number of ARV drugs has been conducted by considering one drug at a time as part of a screening strategy. That is, studies have either restricted analysis to a single drug or drug class, or analyzed exposure to one drug or drug class at a time, and repeated the analysis for each drug and/or drug class (Tuomala et al., 2002; Cotter et al., 2006; Grosch-Woerner et al., 2008; Sibiude et al., 2012; Watts et al., 2013; Koss et al., 2014; Bisio et al., 2015; Perry et al., 2016; Vannappagari et al., 2016; Williams et al., 2016). Such analyses fail to adjust for exposure to other ARV drugs, and thus could be confounded by other ARV use. On the other

hand, with so many different ARV exposures, it can become prohibitive to include all exposures at once in the statistical models ordinarily used.

As an alternative to these conventional approaches, hierarchical modeling has been advocated to address the multiple-exposure issues inherent to many epidemiologic investigations (Greenland, 1992; Greenland, 1993; Witte et al., 1994). It has been used in areas such as nutrition, occupational health and genetics (Greenland, 1992; Witte et al., 1994; Witte and Greenland, 1996; Witte et al., 2000; Greenland, 1997; Aragaki et al., 2003; Conti and Witte, 2003; Hung et al., 2008; Capanu et al., 2008; Capanu and Begg, 2011; Brenner et al., 2013). Hierarchical models have also previously been used in evaluating outcomes among HIV-infected adults, but have not been utilized in the context of addressing safety of ARV use during pregnancy (Young et al., 2009; Wang et al., 2013; Young et al., 2016).

In this paper, we investigate a hierarchical model safety screening approach that includes first-stage effects for each drug class (nucleoside reverse transcriptase inhibitors (NRTI), non-nucleoside reverse transcriptase inhibitors (NNRTI), and protease inhibitors (PI)), and second-stage effects for individual drugs. In essence, this model assumes that the effect of each drug is the summation of the (fixed) effect of its drug class and a residual effect specific to the individual drug. The effect for drugs less commonly used will be pulled toward the "mean" effect averaged over other, more common drugs from its same drug class. We would thus expect the hierarchical modeling method to perform well when drugs from the same drug class do indeed have similar effects on the outcome of interest.

The assumption of a similar effect for drugs within the same drug class can be justified by the fact that each class of antiretroviral medications has a different mechanism of action. NRTIs are analogs of naturally-occurring deoxynucleotides and terminate DNA chain formation

(Kalkut, 2005; Cihlar and Ray, 2010). NNRTIs bind to the HIV reverse transcriptase enzyme and cause a structural change that impairs further DNA synthesis (Kalkut, 2005; De Bethune, 2010). PIs prevent the processing of viral proteins into their functional form, such that release of active virus particles is inhibited (Kalkut, 2005; Wensin et al., 2010). As a result of their mechanism of action, PIs as a class have been linked to increased rates of dyslipidemia in both children and adults with HIV infection (Stein, 2003; Tassiopoulos et al., 2008), and have also been associated with increased rates of preterm birth (Mesfin et al., 2016; Watts et al., 2013), particularly when taken by HIV-infected women early in pregnancy (Uthman et al., 2017). In contrast, NRTIs have been linked to potential mitochondrial dysfunction and lactic acidosis based on evidence from both animal and human studies (Cote et al., 2002). While their common mechanism of action supports an assumption that drugs within a class would behave similarly, and some studies have documented similar rates of outcomes (Perry et al., 2016), there are also specific individual drugs which may confer increased or decreased risk as compared to others within the same class (CDC, 2006; Smith et al., 2016; Abers et al., 2014). For example, the drug efavirenz (EFV) has been more commonly associated with psychiatric adverse effects than other drugs within the NNRTI class (Abers et al., 2014).

Given a plausible biological justification, the hierarchical modeling approach thus seems appealing. However, while a limited number of prior applications have utilized this approach, there is little information on how well this method will perform under various possible scenarios reflecting ARV drug effects. For example, this approach may not perform well when drugs from the same class do not behave similarly. Furthermore, previous research studies utilizing this approach considered multiple continuous exposures with considerably more variability than observed within our context (Witte and Greenland, 1996; Witte et al., 2000). Thus, examination

of whether the hierarchical modeling approach is advantageous within the context of multiple binary exposures with many zero counts is warranted. Given the lack of prior knowledge regarding expected effects in these types of screening studies, we sought to quantify how much is gained by using the hierarchical model when the drug class assumption is correct, and also how much is lost by using the hierarchical model when the drug class assumption contradicts the true underlying data mechanism.

In Section 1.2, we detail the three screening approaches to be compared, and consider the analytical bias of the separate models approach and the hierarchical approach. In Section 1.3, we present a simulation study conducted to compare the conventional approaches and the hierarchical modeling approach under various true exposure-outcome scenarios in the context of screening the safety of ARV exposures during pregnancy. In Section 1.4, we illustrate the hierarchical modeling approach using data from the Surveillance Monitoring of ART Toxicities (SMARTT) study within the Pediatric HIV/AIDS Cohort Network Study (PHACS). In Section 1.5, we conclude with a discussion of the relative merits and limitations of the hierarchical approach for safety screening, and avenues of further research.

1.2 Methods

1.2.1 Models

We consider the setting of an observational cohort study with N participants for whom we have information on ARV exposures during pregnancy and perinatal outcome data. We let $\mathbf{y}$ be an N by 1 outcome vector, indicating a perinatal or infant outcome. We let $\mathbf{X}$ be an N by m matrix of zeroes and ones indicating the exposure history (no/yes) during pregnancy of each participant to m individual ARVs under investigation, and we let $\mathbf{X_j}$ be the N by 1 subvector of $\mathbf{X}$ indicating

the exposure history for the j[th] ARV (j=1,2,..*m*). Lastly, we let $\mathbf{1_N}$ be an N by 1 vector of ones and **W** be an *N by q* matrix of *q* potential confounding variables. Let g(·) denote the link function for a generalized linear model. In particular, we investigate the identity link (g(E(**y**)) = E(**y**)) for continuous outcomes and the logit link (g(E(**y**)) = logit{E(**y**)}) for binary outcomes.

The standard, separate regression models approach involves running *m* models, where each model includes one ARV drug:

$$g\left(\boldsymbol{E}(\boldsymbol{y}|X_j, \boldsymbol{W})\right) = \alpha^S \mathbf{1_N} + \boldsymbol{X_j}\beta_j^* + \boldsymbol{W}\boldsymbol{\gamma_j^*},\ j = 1,2,\dots,m \qquad (1)$$

In Equation (1), α^S represents the mean outcome (under the identity link) or the log odds of the outcome (under the logit link) among those unexposed to the j[th] ARV and for which all covariates in **W** equal zero. The β_j^* represent the mean difference in outcome (under the identity link) or the difference in log odds of the outcome (under the logit link) between women exposed and unexposed to the j[th] ARV after adjusting for the covariates in **W**. The $\boldsymbol{\gamma_j^*}$ is a vector indicating the mean differences in outcome (under the identity link) or the differences in log odds of the outcome (under the logit link) for a one unit increase in the covariates, when adjusting for the j[th] ARV.

The full fixed effect regression model involves running one model with all *m* ARVs included at once:

$$g(\boldsymbol{E}(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{W})) = \alpha^F \mathbf{1_N} + \boldsymbol{X}\boldsymbol{\beta^F} + \boldsymbol{W}\boldsymbol{\gamma^F} \qquad (2)$$

In Equation (2), α^F represents the mean outcome (under the identity link) or the log odds of the outcome (under the logit link) among those unexposed to all *m* ARVs and for which all covariates in **W** equal zero. The $\boldsymbol{\beta^F}$ vector represents the mean differences (or differences in log odds) in outcome under the identity link (or logit link) between women exposed and unexposed to each ARV after adjusting for the other *m* -1 ARVs and the covariates in **W**. The $\boldsymbol{\gamma^F}$ is a vector

indicating the mean differences in outcome (under the identity link) or the differences in log odds of the outcome (under the logit link) for a one unit increase in the covariates, when adjusting for all *m* ARVs.

The hierarchical model adds a prior distribution to the $\boldsymbol{\beta^F}$ coefficients in (2), such that

$$\boldsymbol{\beta^H} = \boldsymbol{Z\pi} + \boldsymbol{\delta},$$

$$\boldsymbol{\delta} \sim N_m(\mathbf{0}, \tau^2 \boldsymbol{I_m}) \quad (3)$$

So, $\boldsymbol{\beta^H} \sim N_m(\boldsymbol{Z\pi}, \tau^2 \boldsymbol{I_m})$, where **Z** is an *m by p* matrix indicating drug class membership when the *m* individual drugs under investigation are from *p* different drug classes, and $\boldsymbol{\pi}$ is a *p by 1* vector of the *p* fixed, drug class-specific mean effects. For example, with *m*=14 drugs from *p*=3 drug classes, **Z** may look like:

	NRTI	NNRTI	PI
Abacavir (ABC)	1	0	0
Emtricitabine (FTC)	1	0	0
Tenofovir (TDF)	1	0	0
Zidovudine (ZDV)	1	0	0
Lamivudine (3TC)	1	0	0
Efavirenz (EFV)	0	1	0
Etravirine (ETR)	0	1	0
Nevirapine (NVP)	0	1	0
Rilpivirine (RPV)	0	1	0
Atazanavir (ATV)	0	0	1
Darunavir (DRV)	0	0	1
Fosamprenavir (FPV)	0	0	1
Ritonavir-boosted Lopinavir (LPV/r)	0	0	1
Nelfinavir (NFV)	0	0	1

$\boldsymbol{\delta}$ is an *m by 1* vector of residual effects for each individual drug, and the elements of $\boldsymbol{\delta}$ are assumed to be independent normal random variables with mean 0 and variance τ^2. The hierarchical model thus becomes:

$$g(\boldsymbol{E}(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W},\boldsymbol{\delta})) = \alpha + \boldsymbol{X}(\boldsymbol{Z}\boldsymbol{\pi} + \boldsymbol{\delta}) + \boldsymbol{W}\boldsymbol{\gamma} = \alpha\mathbf{1}_N + \boldsymbol{X}\boldsymbol{Z}\boldsymbol{\pi} + \boldsymbol{X}\boldsymbol{\delta} + \boldsymbol{W}\boldsymbol{\gamma},$$

$$\boldsymbol{\delta} \sim N_m(\mathbf{0}, \tau^2\boldsymbol{I_m}) \quad (4)$$

From the formulation in (4), we can see that $\boldsymbol{XZ}$ is an *N by p* matrix indicating the *number* of drugs from each drug class that each participant was exposed to during pregnancy. The elements in $\boldsymbol{\pi}$ represent the effect on the outcome of each additional drug from a particular drug class that a woman is exposed to during pregnancy, conditional on the individual drugs taken and covariates in **W**. The elements of $\boldsymbol{\delta}$ are the residual effects on the outcome for a particular drug above and beyond the effects attributed to its drug class. The α parameter represents the mean outcome (under the identity link) or the log odds of the outcome (under the logit link) among those unexposed to all *m* ARVs and for which all covariates in **W** equal zero; and $\boldsymbol{\gamma}$ is a vector of the covariate effects conditional on exposure to drug classes and individual drugs.

The variance of the random effects (τ^2) controls the degree of shrinkage of the β^H's to their drug class mean. Smaller values of τ^2 will result in more shrinkage to the drug class mean, with the hierarchical model reducing to a model with just fixed effects for drug class when τ^2=0. Larger values of τ^2 correspond to less shrinkage to the drug class mean, and the hierarchical model becomes equivalent to the ordinary full regression model when $\tau^2 = \infty$.

1.2.2 Brief bias considerations under the linear model

As mentioned earlier, we would expect the hierarchical modeling method to perform well when drugs from the same drug class have similar effects on the outcome of interest. However, often

there is little prior knowledge regarding the effects of ARV exposures on reproductive and perinatal outcomes, and the relative advantages of the hierarchical approach when only a subset of ARV drugs have an effect requires evaluation. Suppose the true underlying data generating mechanism is that only one drug, X_1, has an effect on a continuous outcome Y in the following form:

$$y_i = \alpha^* + X_1\beta_1^* + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Under the separate models approach, the maximum likelihood estimate (MLE) for β_1^* will be unbiased and consistent when fitting drug 1, i.e. the correct model. However, MLE estimates for the β_j^* from the other m -1 models will be biased due to uncontrolled confounding by X_1. In particular, it can be shown that the expected value of $\hat{\beta}_j^*$ has the form

$$E\left[\hat{\beta}_j^*\right] = \xi_{1j}\beta_1^*, \quad j = 2, 3, \ldots, m$$

where ξ_{1j} is the difference in probability of receiving drug X_1 between women exposed and unexposed to drug X_j, i.e:

$$E\left[X_1 \middle| X_j\right] = P\left(X_1 = 1 \middle| X_j\right) = \xi_{0j} + \xi_{1j}X_j, \quad j = 2, 3, \ldots, m$$

Thus, the MLE estimators from a separate models approach will be biased for the true null effect ($\beta_j^* = 0$). As the magnitude of the effect of X_1 on Y (β_1^*) increases, and as the correlation between exposure to drug X_1 and drug X_j (ξ_{1j}) increases, the bias in $\hat{\beta}_j^*$ also increases. Furthermore, increasing the sample size only exacerbates the problem, as the separate models approach will show increasing certainty (smaller standard errors) around an incorrect effect estimate in m-1 of the models.

Often researchers adjust for potential confounders between the drug exposures and the outcome. However, the confounded effect estimate of X_j will remain unless the model controls for all covariates $\boldsymbol{W}^*$ that determine prescribing patterns by physicians such that $\xi_{1j}^* = 0$ under

$E\left[X_1 \middle| X_j, \boldsymbol{W}^*\right] = P\left(X_1 = 1 \middle| X_j\right) = \xi_{0j}^* + \xi_{1j}^* X_j + \boldsymbol{W}^* \boldsymbol{\theta}$. Given the differences in prescribing patterns across hospitals and physicians, it seems unlikely one could fully account for $\boldsymbol{W}^*$.

Under the hierarchical modeling approach, the estimated drug-specific effects are also biased, but the bias decreases as the sample size increases. Greenland (1993) and Greenland (1997) noted that $\widehat{\boldsymbol{\beta}}^{\boldsymbol{H}} = \mathbf{BZ}\widehat{\boldsymbol{\pi}} + \mathbf{(I-B)}\,\widehat{\boldsymbol{\beta}}^{\boldsymbol{F}}$, where $\mathbf{B}=(\boldsymbol{V}^* + \tau^2 \boldsymbol{I_m})^{-1}\boldsymbol{V}^*$, and $\mathbf{V^*}$ is the covariance matrix of $\widehat{\boldsymbol{\beta}}^{\boldsymbol{F}}$. For a given τ^2, as $\mathbf{V^*} \rightarrow \mathbf{0}$ with increasing sample size, $\widehat{\boldsymbol{\beta}}^{\boldsymbol{F}}$ is given more weight and $\widehat{\boldsymbol{\beta}}^{\boldsymbol{H}}$ is a consistent estimator for the true parameters of all *m* drugs. That is, as $N \rightarrow \infty$, $\hat{\beta}_1^H \rightarrow \beta_1^*$ and $\hat{\beta}_j^H \rightarrow 0$ for $j=2,3,\ldots m$. Asymptotic properties, however, may not be reasonable approximations for estimators at the sample sizes commonly utilized for studies assessing ARV exposures and reproductive outcomes. In this paper, we will consider the bias under both methods at realistic sample sizes to assess finite-sample properties and further consider the bias under a binary outcome with generalized linear models.

1.3 Simulation Study

A simulation study was performed to investigate the operating characteristics of the three different approaches under various outcome scenarios. The first approach involved separate univariate regression models for each drug (Equation 1); the second approach was the full ordinary regression model with all drugs included at once (Equation 2); and the third approach was the hierarchical model (Equation 4). We used a semi-Bayes approach for fitting the hierarchical model by specifying *a priori* the variance in the random effects (τ^2), as advocated in prior studies using this approach (Greenland, 1993; Witte et al., 2000; Wang et al., 2013; Young et al., 2016; Greenland 2000). An empirical Bayes approach (estimating τ^2 from the data) was also considered, but τ^2 was consistently estimated to be zero, which reduces the model to having

only fixed effects for drug class and is not helpful in making drug-specific conclusions. We considered a binary outcome (preterm birth) and a continuous outcome (Bayley-III score of the infant at 12 months). For each outcome, we considered various true exposure-outcome relationships, including no true effects, a subtle effect of all drugs within one drug class, a moderate effect of only one individual drug, and moderate effects of two drugs from the same class, but in opposite directions. Table 1.1 provides the specific models under which data were simulated for each scenario.

A number of statistical properties were evaluated, including the percent of models that converged (for the binary outcome), the percent of false discoveries, the power to detect true effects, the bias in estimated effects for each exposure, the standard error in estimated effects for each exposure, and the observed coverage of 95% confidence intervals for the effect for each exposure.

SAS 9.4 (SAS Institute Inc., Cary, North Carolina) was used for all simulations and applied data analysis. The SAS-provided GLIMMIX macro (http://support.sas.com/techsup/notes/v8/25/030.html) was used to implement the hierarchical modeling method for the binary outcome (Witte et al., 2000). Note that the GLIMMIX procedure does not yield estimates of the covariances between fixed and random effects, and thus cannot be used for this approach. The MIXED procedure was used to implement the hierarchical modeling method for the continuous outcome (programs are available by request to the author).

Table 1.1. A summary of the true exposure-outcome relationship scenarios considered in the simulation studies.

Scenario	Binary outcome: Preterm birth (<37 weeks gestational age at delivery)[a]	Continuous outcome: Standardized Bayley-III score of infant at 12 months
(i) No effects	$P(Y_1) = 0.12$	$E(Y_2) = 0$
(ii) A class of drugs has a subtle effect	$P(Y_1) = 0.12 + 0.04$*PI ($OR_{PI} = 1.40$)	$E(Y_2) = 0 - 0.3$*PI
(iii) One ARV drug has a moderate effect		
(a) more common ARV drug (>15% exposure)	$P(Y_1) = 0.12 + 0.09$*LPV/r ($OR_{LPV/r} = 1.95$)	$E(Y_2) = 0 - 0.5$*LPV/r
(b) less common ARV drug (5–15% exposure)	$P(Y_1) = 0.12 + 0.09$*ABC ($OR_{ABC} = 1.95$)	$E(Y_2) = 0 - 0.5$*ABC
(c) rarely used ARV drug (<5% exposure)	$P(Y_1) = 0.12 + 0.09$*EFV ($OR_{EFV} = 1.95$)	$E(Y_2) = 0 - 0.5$*EFV
(iv) Two drugs from the same drug class have moderate effects, but in opposite directions	$P(Y_1) = 0.12 + 0.09$* LPV/r $- 0.05$*DRV ($OR_{LPV/r} = 1.95$, $OR_{DRV} = 0.55$)	$E(Y_2) = 0 - 0.5$*LPV/r $+ 0.5$*DRV

ABC: abacavir; DRV: darunavir; E: expected value; EFV: efavirenz; LPV/r: ritonavir-boosted lopinavir; OR_{ABC}: odds ratio comparing ABC-exposed to ABC-unexposed; OR_{DRV}: odds ratio comparing DRV-exposed to DRV-unexposed: OR_{EFV}: odds ratio comparing EFV-exposed to EFV-unexposed; $OR_{LPV/r}$: odds ratio comparing LPV/r-exposed to LPV/r-unexposed; OR_{PI}: odds ratio comparing PI-exposed to PI-unexposed; P: probability; PI: protease inhibitor; Y_1: preterm birth; Y_2: standardized Bayley-III cognitive score.

[a]The corresponding logistic models are: (i) $\text{logit}(P(Y_1)) = -1.9924$; (ii) $\text{logit}(P(Y_1)) = -1.9924 + 0.3365$*PI; (iii) $\text{logit}(P(Y_1)) = -1.9924 + 0.6678$*$X_j$ (where X_j indicates LPV/r, ABC, or EFV); and (iv) $\text{logit}(P(Y_1)) = -1.9924 + 0.6678$*LPV/r $- 0.5978$*DRV. Note that LPV/r and DRV are mutually exclusive (women are never exposed to both drugs simultaneously).

1.3.1 Exposure assignment

We used data from the SMARTT study to inform the ARV exposure distributions within the simulation study. The SMARTT study is a large cohort study with data on HIV-uninfected children born to HIV-infected women since 1995 to the present. Patterns in ARV use during pregnancy have changed dramatically over these years, but HIV-infected women typically receive a combination regimen during pregnancy consisting of a two-NRTI backbone plus either a PI or an NNRTI (Griner et al., 2011). We are specifically interested in monitoring the safety of current combination regimens, and thus used the observed distribution of regimens reported in SMARTT between 2010 and 2015 to inform the exposure distribution. In particular, regimens were assigned via a multinomial distribution with 107 categories (for the 107 different observed regimens over this time period), with each category having the same probability (ranging between 0.0008-0.2264) as observed in the SMARTT cohort. Exposures to 14 individual drugs and three drug classes were then derived from the assigned regimen. Specifically, five NRTIs, four NNRTIs, and five PIs were included in the simulation analysis, as shown in the **Z** matrix in Section 1.2.1.

1.3.2 Outcome assignment

We acknowledge that it is improbable the hierarchical model being fit reflects the true underlying outcome mechanism. Rather, our interest lies in whether a hierarchical model can be a useful screening approach despite violations to its underlying assumptions. Consequently, outcomes were assigned randomly via the Bernoulli distribution (for preterm birth) or the standard Normal distribution (for standardized Bayley-III score) under simple models based on exposure and outcome scenario (see Table 1.1). Three thousand simulated datasets were created

in this way. The main simulations were conducted with a sample size of 1,000. Additional simulations were conducted with sample sizes of 500, 3,000, and 5,000.

For the binary outcome, the hierarchical model was fit specifying a τ^2 value of 0.125, which corresponds to 95% of the *residual* effects of a particular ARV drug (above and beyond the effects of its drug class) lying between odds ratios of ½ and 2 ($[e^{-1.96/\sqrt{8}}, e^{1.96/\sqrt{8}}]$). We also considered τ^2 values of 0.36 and 0.64, which are equivalent to allowing residual effects to fall within an expanded 10-fold and 25-fold range, respectively, but simulation results presented for the binary outcome are for τ^2 =0.125 (Greenland, 1993). For the continuous outcome, the hierarchical model was fit specifying a τ^2 value of 0.26, corresponding to 95% of the *residual* effects of a particular drug falling within one standard deviation. Additional analyses considered values of 1.04 and 2.34, equivalent to allowing residual effects to fall within two and three standard deviations, respectively.

1.3.3 Simulation results

For the binary outcome, convergence of the model was a sizeable problem with the full model but a minimal issue with the hierarchical model. At a sample size of 1,000, all of the hierarchical models converged under each outcome scenario, whereas the full logistic model failed to converge in 14- 22% of simulations, depending on the outcome scenario. With *N*=500, the full model failed to converge in over 75% of the simulations, while the hierarchical model failed to converge in 0.1% of simulations. The separate model approach converged for all 13 models over 95% of the time; however, results for rare exposures were sometimes nonsensical, with standard errors exceeding 500. For instance, the simple logistic model failed to yield interpretable results

for efavirenz (EFV) in up to 24% of the simulations at N=1,000 and in up to 40% of the simulations at N=500.

The hierarchical model outperformed both the full model and the separate model approaches in terms of false discoveries, regardless of outcome type and outcome scenario (Figure 1). With a binary outcome, the hierarchical model had no false discoveries over 80% of the time. The full model had no false discoveries for 64% (under scenario (i)) to 74% (under scenario (ii)) of simulations. The separate model approach had false discovery rates comparable to the full model approach under scenarios (i) and (ii), but did quite poorly under scenarios (iii.a) and (iv). Notably, under the latter two scenarios, the standard approach had at least one false discovery in over 70% of the simulations, and four or more false discoveries (of twelve truly null effects) in 40% of simulations under scenario (iv).

For the continuous outcome, false discovery rates were consistently higher than observed for the binary outcome, though the hierarchical model maintained noticeably lower rates than the other two methods (Figure 1.1). Under scenarios (iii.a) and (iv), the separate models method identified one or more false discoveries in over 99% of the simulations, and four or more false discoveries in over 90% of the simulations.

Detection of true effects is irrelevant to scenario (i). With N=1,000, the true effects of the five PIs under a common drug class assumption (scenario (ii)) were detected most often by the hierarchical model for both outcome types (Figure 1.2). This result was to be expected because the hierarchical model assumes drugs from the same class behave similarly, which corresponds to the true underlying data mechanism in this scenario. For the remaining scenarios, detection of true effects differed depending on outcome type. With a binary outcome, the hierarchical model performed similarly to the full fixed effect model but substantially worse than the separate

models method in detecting the true effects of the ARVs in scenarios (iii.a), (iii.b), (iii.c), and (iv). This result also was to be expected given that the hierarchical model assumes similar effects for drugs from the same class, which is not correct in scenarios (iii) and (iv). Interestingly, however, under the continuous outcome, all three methods detected the true effects of the ARVs almost 100% of the time in scenarios (iii.a), (iii.b), and (iv). Under scenario (iii.c), the separate models method detected the true effect of efavirenz (EFV) more often than the other two methods, though the differences were not as large as under the binary outcome (Figure 1.2).

The additional simulations showed that as the sample size increases, the hierarchical model continued to detect the true effects of the PIs under scenario (ii) considerably more often than the separate models method, while also continuing to minimize the number of false discoveries. With the continuous outcome, all three methods detected the true effects of the ARVs equally under the other scenarios by N=3,000 (Figure 1.2). With a binary outcome, the hierarchical model detected the true effects about as well as the other methods at N=5,000 for scenarios (iii.a), (iii.b) and (iv), but failed to detect the true effect of efavirenz (EFV) as often as the other methods under scenario (iii.c) even for N=5,000 (Figure 1.2).

Simulation results under scenario (iv) for the bias and standard errors (SE) in estimated coefficients and coverage of 95% confidence intervals (CI) among the three approaches are presented in Table 1.2 for the binary outcome and Table 1.3 for the continuous outcome. Scenario (iv) represents the "worst-case" type scenario for the hierarchical model since the prior being fit (assuming drugs from the same class behave similarly) contradicts the true underlying exposure-outcome relationship. Still, some patterns in these results remain consistent across scenarios (see Tables A1.1-A1.10 in Appendix). First, SEs were consistently largest under the full model. For rare exposures (<5% exposed), the SEs were smallest under the hierarchical

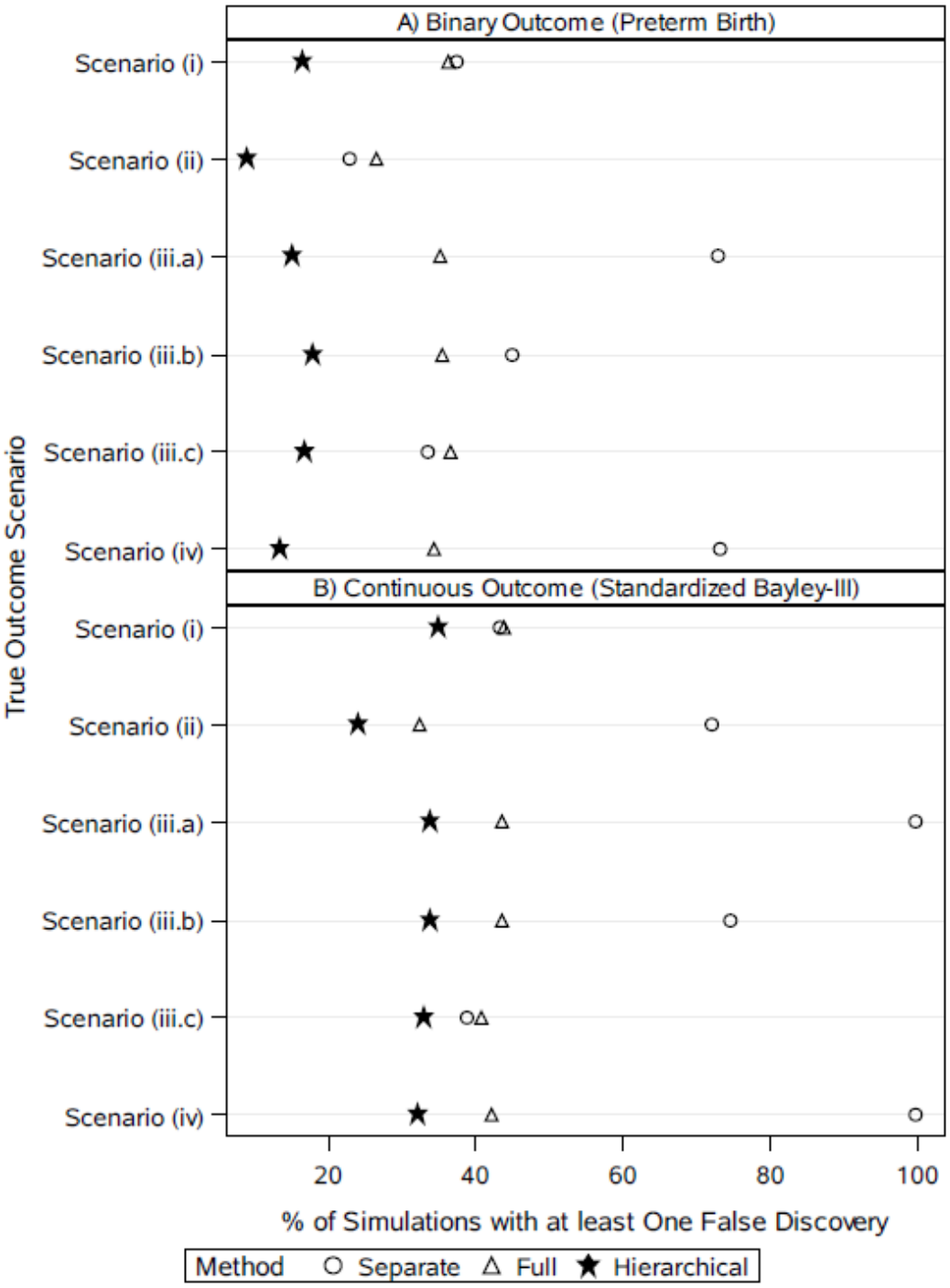

Figure 1.1. The percent of simulations with at least one false discovery at a sample size of 1000 under three statistical approaches and six different true outcome-exposure relationships, by outcome type (a) binary; or (b) continuous. Each scenario considers 14 different antiretroviral drugs. See Table 1.1 for Scenario specifications. Results based on 3,000 simulations.

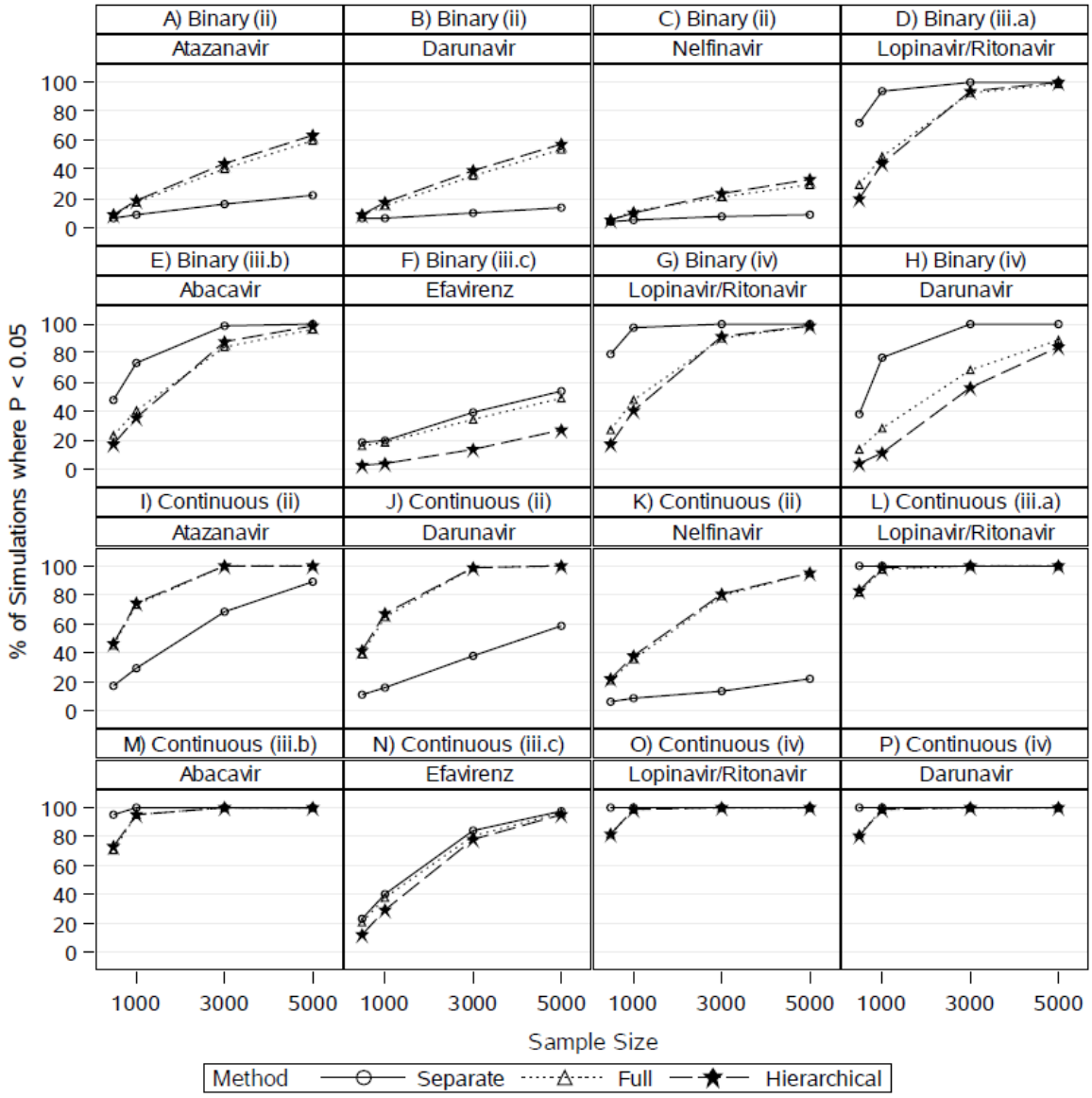

Figure 1.2. The power to detect true effects of antiretroviral (ARV) exposures on preterm birth and standardized Bayley-III score as a function of sample size under three statistical approaches and six different true outcome-exposure relationships. Results are based on 3,000 simulations. Each panel reflect the power to detect the true effect of an ARV drug under a specific scenario as outlined in Table 1.1.

model, but for the more common exposures (>15% exposed), they were smallest under the separate models method. Second, the bias in estimated coefficients tended to be minimized under the hierarchical model, the main exception being for when an uncommon drug was the only drug with a true effect (e.g. abacavir (ABC) in scenario (iii.b) and efavirenz (EFV) in scenario (iii.c)). Third, the nominal coverage rates of the 95% CIs were quite poor for some of the ARVs under the separate models method. The poor coverage rates tended to be for more common drugs that had relatively high bias (due to uncontrolled confounding by other ARV exposures) and relatively small SEs. For example, under scenario (iv), the 95% CI for zidovudine (ZDV) captured its true effect (null) in only 59% of the simulations for the binary outcome (Table 1.2) and in only 1% of the simulations for the continuous outcome (Table 1.3).

Additional simulations were conducted to assess how results may vary for binary outcomes that are much rarer or much more common than the moderate baseline prevalence (0.12) considered in the main simulations. In particular, baseline prevalences of 0.25 and 0.05 were considered. Although power increased for the more common outcome and decreased for the less common outcome, the relative differences across the three approaches remained similar to results from the main simulations and thus results are not shown here.

1.4 Illustrative example

We applied the hierarchical modeling approach to evaluate ARV use and preterm birth in the SMARTT cohort. The SMARTT study has been approved by the research ethics committee at Harvard T.H. Chan School of Public Health and all research sites, and study participants provided written informed consent. The SMARTT cohort has enrolled over 3,000 HIV-infected pregnant women from 22 sites around the United States, as described elsewhere (Watts et al.,

Table 1.2. Bias, efficiency, and coverage of 95% confidence intervals across statistical approaches for the binary outcome under scenario (iv) (two drugs from the same drug class have opposite effects on preterm birth; DRV has a protective effect and LPV/r has a detrimental effect).

		Mean bias in log odds			Mean SE of log odds			Coverage of 95% CI		
% Exposed	Drug	Separate[a]	Full[b]	Hierarchical	Separate[a]	Full[b]	Hierarchical	Separate[a]	Full[b]	Hierarchical
<5%	EFV	0.042	0.237	−0.008	0.862	0.914	0.516	0.98	0.96	0.99
	ETR	−0.287	0.055	−0.035	0.788	0.827	0.492	0.99	0.97	0.99
	NVP	−0.231	−0.062	−0.026	0.669	0.728	0.474	0.98	0.97	0.97
	FPV	−0.101	0.040	0.067	0.792	0.852	0.445	0.98	0.97	0.99
	NFV	−0.268	−0.050	0.021	0.493	0.591	0.402	0.97	0.96	0.98
5–15%	ABC	−0.187	−0.017	−0.040	0.297	0.424	0.297	0.93	0.95	0.97
	RPV	−0.248	−0.039	−0.003	0.436	0.523	0.423	0.96	0.96	0.96
	DRV	−0.306	−0.035	0.296	0.357	0.446	0.350	0.92	0.94	0.89
>15%	3TC	0.369	0.072	0.049	0.185	0.713	0.339	0.49	0.95	1.00
	FTC	−0.357	−0.017	−0.002	0.186	0.919	0.345	0.53	0.97	0.99
	TDF	−0.356	0.116	−0.002	0.185	0.928	0.342	0.52	0.96	0.99
	ZDV	0.386	0.022	0.060	0.184	0.474	0.297	0.44	0.94	0.98
	ATV	−0.234	0.011	0.048	0.219	0.358	0.319	0.84	0.95	0.96
	LPV/r	0.093	0.031	−0.118	0.189	0.363	0.312	0.92	0.95	0.94

ABC: abacavir; ATV: atazanavir; CI: confidence interval; DRV: darunavir; EFV: efavirenz; ETR: etravirine; FPV: fosamprenavir; FTC: emtricitabine; LPV/r: ritonavir-boosted lopinavir; NFV: nelfinavir; NVP: nevirapine; RPV: rilpivirine; SE: standard error; TDF: tenofovir; ZDV: zidovudine; 3TC: lamivudine.

[a]21.8% of the EFV models, 13.9% of the ETR models, 13.5% of the FPV models, 0.3% of the NFV models, 5.0% of the NVP models, and 0.1% of the RPV models did not converge or yielded unreasonable SEs. Results presented exclude these models.

[b]18.8% of the full models did not converge. Results presented exclude these models.

Note: Results are from 3000 simulations each with sample size 1000.

Table 1.3. Bias, efficiency, and coverage of 95% confidence intervals across statistical approaches for the continuous outcome under scenario (iv) (two drugs from the same drug class have opposite effects on Bayley-III score; DRV has a protective effect and LPV/r has a detrimental effect).

		Mean bias in difference			Mean SE of difference			Coverage of 95% CI		
% Exposed	Drug	Separate	Full	Hierarchical	Separate	Full	Hierarchical	Separate	Full	Hierarchical
<5%	EFV	0.081	0.007	0.004	0.313	0.314	0.280	0.95	0.95	0.96
	ETR	0.254	0.002	0.006	0.250	0.248	0.231	0.83	0.95	0.96
	NVP	0.073	−0.002	0.001	0.214	0.221	0.210	0.94	0.94	0.95
	FPV	0.042	−0.007	−0.008	0.269	0.273	0.249	0.95	0.95	0.96
	NFV	0.081	−0.006	0.001	0.157	0.182	0.176	0.93	0.95	0.96
5–15%	ABC	0.098	−0.003	0.005	0.101	0.140	0.132	0.84	0.96	0.97
	RPV	0.087	−0.005	−0.006	0.141	0.161	0.157	0.91	0.95	0.96
	DRV	0.168	−0.000	−0.013	0.093	0.123	0.121	0.56	0.95	0.95
>15%	3TC	−0.265	0.002	−0.003	0.066	0.224	0.201	0.02	0.96	0.97
	FTC	0.251	0.003	0.005	0.066	0.309	0.244	0.03	0.95	0.98
	TDF	0.250	−0.003	−0.001	0.066	0.310	0.243	0.03	0.94	0.98
	ZDV	−0.284	0.001	−0.002	0.066	0.154	0.143	0.01	0.95	0.97
	ATV	0.102	0.001	−0.001	0.076	0.114	0.112	0.73	0.95	0.95
	LPV/r	−0.100	−0.003	0.010	0.071	0.123	0.120	0.71	0.95	0.95

ABC: abacavir; ATV: atazanavir; CI: confidence interval; DRV: darunavir; EFV: efavirenz; ETR: etravirine; FPV: fosamprenavir; FTC: emtricitabine; LPV/r: ritonavir-boosted lopinavir; NFV: nelfinavir; NVP: nevirapine; RPV: rilpivirine; SE: standard error; TDF: tenofovir; ZDV: zidovudine; 3TC: lamivudine.
Note: Results are from 3000 simulations each with sample size 1000.

2013). Consistent with prior analyses, we controlled for birth cohort (1995-2004, 2005-2009, 2010-2012, and 2013-2015), annual income <$20,000, and black race (Watts et al., 2013).

Our analysis included 2,660 singleton pregnancies with ARV exposures and preterm birth outcomes available. The majority of women (71%) received only one ARV regimen during their pregnancy. For this analysis, we classified the maternal ARV regimen as that taken for the longest duration during pregnancy, and considered a woman exposed to a particular drug if that drug was included in her most common regimen. We assessed 18 individual drugs, including seven NRTIs, four NNRTIs, and seven PIs.

Table 1.4 presents odds ratios (OR) and 95% CIs from the hierarchical model under three different values of τ^2 and from the full logistic model (equivalent to the hierarchical model at $\tau^2 = \infty$). Consistent with results from the simulation study, as τ^2 increased, the CIs tended to widen, with the CIs widest under the full logistic model. The shrinkage effect of the hierarchical model can be observed for rarely used ARVs, for which estimated ORs in the hierarchical model are further from their estimated ORs under the full model (i.e. they are being pulled more toward their drug class mean effect), whereas the estimated ORs for common drugs were more similar. For example, the estimated OR for the least common PI (indinavir (IDV)) was 1.24 (95% CI: 0.66, 2.31) in the hierarchical model with τ^2 = 0.125 and 1.51 (95% CI: 0.61, 3.73) in the full model. In comparison, the estimated ORs from those models for the most common PI (ritonavir-boosted lopinavir (LPV/r)) were 1.51 (95% CI: 1.10, 2.06) and 1.50 (95% CI: 1.08, 2.09), respectively. In addition, as τ^2 increases, the estimated ORs from the hierarchical model get closer to the estimated ORs from the full model. For example, for indinavir (IDV), the estimated ORs are 1.24 (95% CI: 0.66, 2.31), 1.34 (95% CI: 0.62, 2.89), and 1.39 (95% CI: 0.61, 3.17) under τ^2 values of 0.125, 0.36, and 0.64, respectively.

Table 1.4 Odds ratios (OR) and 95% confidence intervals (CI) for individual antiretroviral drug exposures and preterm birth from the Surveillance Monitoring for ART Toxicities Study cohort of 2668 singleton pregnancies between 1995 and 2015.

			Hierarchical model						Full logistic model	
			$\tau^2 = 0.125$		$\tau^2 = 0.36$		$\tau^2 = 0.64$		$(\tau^2 = \infty)$	
Drug Class	Drug	% Exposed	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
NRTI	ABC	16.2	0.96	0.71, 1.30	0.98	0.71, 1.35	0.98	0.71, 1.37	0.99	0.71, 1.38
	DDI	1.9	0.79	0.46, 1.38	0.79	0.41, 1.53	0.79	0.39, 1.60	0.80	0.37, 1.72
	D4T	2.6	0.74	0.44, 1.25	0.69	0.37, 1.28	0.67	0.35, 1.28	0.63	0.31, 1.28
	FTC	26.8	0.75	0.47, 1.19	0.74	0.43, 1.29	0.74	0.41, 1.34	0.74	0.39, 1.43
	TDF	29.3	0.79	0.52, 1.19	0.81	0.50, 1.31	0.82	0.49, 1.36	0.83	0.47, 1.47
	ZDV	63.1	0.67	0.48, 0.95	0.64	0.44, 0.94	0.63	0.42, 0.93	0.61	0.40, 0.92
	3TC	64.6	0.83	0.57, 1.19	0.85	0.56, 1.29	0.86	0.55, 1.33	0.88	0.55, 1.41
NNRTI	EFV	0.9	0.98	0.47, 2.01	0.83	0.34, 2.05	0.76	0.28, 2.02	0.58	0.17, 2.03
	ETR	0.8	1.58	0.77, 3.23	2.04	0.88, 4.77	2.32	0.94, 5.68	3.01	1.13, 7.98
	NVP	6.4	1.09	0.70, 1.70	1.07	0.68, 1.71	1.07	0.67, 1.71	1.05	0.65, 1.71
	RPV	2.3	1.08	0.58, 2.00	1.02	0.51, 2.04	0.99	0.48, 2.04	0.93	0.42, 2.04
PI	ATV	15.3	0.84	0.58, 1.24	0.80	0.53, 1.19	0.78	0.52, 1.17	0.75	0.49, 1.14
	DRV	6.3	0.88	0.56, 1.38	0.81	0.50, 1.31	0.78	0.47, 1.28	0.72	0.42, 1.22
	FPV	2.1	0.88	0.49, 1.59	0.76	0.38, 1.54	0.71	0.33, 1.50	0.61	0.26, 1.45
	IDV	1.2	1.24	0.66, 2.31	1.34	0.62, 2.89	1.39	0.61, 3.17	1.51	0.61, 3.73
	LPV/r	26.8	1.51	1.10, 2.06	1.52	1.10, 2.10	1.52	1.09, 2.10	1.50	1.08, 2.09
	NFV	16.3	0.95	0.67, 1.35	0.94	0.65, 1.36	0.94	0.65, 1.36	0.93	0.63, 1.36
	SQV	2.1	1.71	1.01, 2.89	2.01	1.12, 3.60	2.12	1.17, 3.86	2.31	1.25, 4.27

ABC: abacavir; ATV: atazanavir; CI: confidence interval; DDI: didanosine; DRV: darunavir; D4T: stavudine; EFV: efavirenz; ETR: etravirine; FPV: fosamprenavir; FTC: emtricitabine; IDV: indinavir; LPV/r: ritonavir-boosted lopinavir; NFV: nelfinavir; NNRTI: non-nucleoside reverse transcriptase inhibitor; NRTI: nucleoside reverse transcriptase inhibitor; NVP: nevirapine; OR: odds ratio; PI: protease inhibitor; RPV: rilpivirine; SE: standard error; SQV: saquinavir; TDF: tenofovir; ZDV: zidovudine; 3TC: lamivudine.

Results from the hierarchical model with τ^2=0.125 suggest that further studies should focus on the possible detrimental associations between saquinavir (SQV) and ritonavir-boosted lopinavir (LPV/r) and preterm birth (Table 1.4), as both these drugs have relatively high estimated odds ratios (>1.5) with fairly little variability around the estimates (95% CIs: 1.01, 2.89 and 1.10, 2.06, respectively). The estimated odds ratio for etravirine (ETR) is also relatively high (OR=1.58), but with just 8% of women exposed to etravirine (ETR) in pregnancy, there is much more variability around that estimate (95% CI: 0.77, 3.23), suggesting follow-up on etravirine (ETR) would take lower priority than follow-up on saquinavir (SQV) and ritonavir-boosted lopinavir (LPV/r).

1.5 Discussion

We evaluated how a hierarchical modeling approach to screening ARV use in pregnancy would operate in practice under various conditions. In theory, a hierarchical model offers a compromise between evaluating individual ARV drugs one at a time (which is the current method of choice for assessing the safety of ARV exposures in pregnancy) and fitting a full fixed effect model. It has the benefit of adjusting for other ARV exposures like the full model, but has less convergence problems, smaller standard errors, and more stable estimates than a full fixed effect model approach. However, the hierarchical model groups ARVs from the same drug class together, when there is often little prior knowledge regarding possible effects and the underlying biological mechanisms that ARVs have on perinatal and infant outcomes. If drugs from the same class have disparate effects on an outcome, adopting a hierarchical model approach for ARV safety screening could potentially undermine the screening approach.

In this study, we compared the performance of three different approaches under six different underlying true exposure-outcome relationships. Our results suggest that the hierarchical model that groups ARVs by drug class is almost always advantageous with a large enough sample (e.g. 5,000). It minimizes the number of false negatives under each scenario as compared to both the full and separate models; it is able to detect the true effects substantially better than the separate models method and as well as or slightly better than the full model method when drugs from the same class behave similarly; and is still able to detect true effects similarly to the other methods even when drugs from the same class have opposite effects, except in the case of a binary outcome with a rare exposure.

In reality, however, these types of safety screening studies usually have smaller sample sizes, and the implications of the simulation study for use of the hierarchical model in smaller samples are less straightforward. If we wish to optimize the detection of true effects regardless of the expense in false discovery, then determining which approach to employ may involve taking into account the strength of one's prior belief regarding effects of drugs from the same class, the sample size, and the outcome type (binary or continuous). However, perhaps one of the surprising results from the simulations was just *how* high the false discovery rate can be when evaluating ARV drugs individually, with four or more false discoveries (among 12 drugs) over 90% of the time, and abysmal nominal coverage rates of 95% confidence intervals for some drugs in certain scenarios. Its poor performance in these areas is largely due to biased effect estimates from uncontrolled confounding by other ARV exposures. Power considerations in such settings become irrelevant when there are numerous false signals detected, and as a result evaluating ARVs individually may not allow identification of safety signals to appropriately focus future studies (see Figures A1.1 and A1.2 in Appendix).

We present the hierarchical modeling approach as a screening approach, where little prior knowledge is available regarding possible exposure-outcome relationships. However, if there is evidence of differing effects for drugs belonging to the same class, then the full model may be suggested as a first choice for model fit. Particularly for rare drugs and a binary outcome, the full model has more power to detect the true effects if drugs from the same class do not have similar effects on the outcome; the full model also exhibits less bias in the effect estimates for the drugs with the true effects and better nominal coverage rates for the 95% confidence intervals for the drugs with true effects. Thus, presuming the model converges, the full model has advantages over the hierarchical model when drugs from the same class do not behave similarly on an outcome. Nonetheless, if the full model does not converge, the hierarchical model specified with a large variance for the random effects (τ^2) to allow larger residual effects for individual drugs is an appropriate alternative.

Our simulations and applied data analysis considered drugs from three drug classes (NRTIs, NNRTIs, and PIs). The number of drug classes has expanded in recent years, and as new drugs from new drug classes are made available (e.g. fusion inhibitors, entry inhibitors), some drugs may be the only drug of their drug class. For these drugs, the advantages of the hierarchical model are limited. Drugs unique to their class could still be included in a hierarchical model as fixed effects, but they would not be able to "borrow" information from other drugs in their class. Alternatively, Wang et al grouped rare drugs unique to their class together in an "other" category (Wang et al., 2013). The drug class effect for this "other" group does not have any clinical meaning, but it may still improve the reliability of the estimates for those rare drugs. In particular, based on our simulation results, it may be an advantageous option so long as drugs in the "other" group do not have opposite effects.

We did not consider any interactions between ARVs in this study. Further research is needed to characterize how the hierarchical model performs when interactions are present.

This study highlights the shortcomings – in particular, the inherent bias – of the separate models approach that is currently used to screen the safety of ARVs used during pregnancy. A hierarchical modeling approach can be a superior alternative to the current method, particularly when considering a binary outcome in large samples (N >3,000), a continuous outcome in moderate or large samples (N > 500), and/or when there is prior evidence suggesting drugs from the same class behave similarly on the outcome of interest.

2. Estimating the relative excess risk due to interaction in clustered data settings

The risk difference scale is often of primary interest when evaluating public health impacts of interventions on binary health outcomes, and particularly when considering interaction effects between exposures (Rothman et al, 1980; Rothman, 1998; Rothman et al., 2008; Aschengrau et al., 2014; Vanderweele, 2015). Estimates of additive interaction are more useful than those of multiplicative interaction in order to identify target subpopulations for most effective use of resources (Vanderweele, 2015). Vanderweele (2015) provides a thorough discussion on additive and multiplicative interaction, including examples demonstrating why additive interaction is the more relevant measure for assessing public health relevance (pages 252-253 and section 9.5). Despite the importance of assessing interaction as departure from additivity, models most often used for binary outcomes implicitly measure interaction on the multiplicative scale. Very few studies have incorporated additive interaction into presentation of findings, although recommendations support reporting both measures (Vanderweele, 2015; Knol et al., 2009; Knol et al., 2012).

One measure to assess additive interaction from multiplicative models is the relative excess risk due to interaction (RERI). The RERI measure has been applied in many contexts, including hypertension research (Timpka et al., 2017; Jian et al., 2017), cardiology (Meng et al., 2015; Vart et al., 2015; Zhang et al., 2015; Gustavsson et al., 2016; Crump et al., 2017; Hagihara et al., 2017), oncology (Menvielle et al., 2016; Oh et al., 2016; Simons et al., 2016; White et al., 2017), and genetics (Gustavsson et al., 2016; Simons et al., 2016; Wang et al., 2017). However,

one limitation of current approaches is that clustering in data has rarely been considered (Chen et al., 2006; Aanerud et al., 2015; Jabbarpoor et al., 2016; Mao et al., 2017; Jabbarpoor et al., 2017). In practice, data are often clustered such that outcomes among observations within the same cluster are not independent. Clustering in epidemiological research arises in many forms, including clustering of patients by clinical center or health care provider (Bermedo-Carrasco et al., 2015; Dupont et al., 2017; Raifman et al., 2017; Goyette et al., 2018), clustering of individuals by spatial location (Gemperli et al., 2004; Kloog et al., 2015; Lin et al., 2017), repeated measures taken on the same individual (Hajat et al., 2015; Tsai et al., 2015; Chiu et al., 2018; Madden et al., 2018), and meta-analyses (Cook et al., 2005; White et al., 2008).

In an effort to further encourage the reporting of additive interaction measures for binary outcomes, we evaluate the RERI metric in both population-averaged models and cluster-conditional models in clustered data settings, with a particular focus on more common outcomes. We present results from simulation studies across a range of outcome prevalences to assess the statistical operating characteristics of various approaches. We apply the methods to an observational study of adverse birth outcomes in mothers with HIV infection, in which enrolled mothers were clustered within clinical research sites.

2.2 Approaches for estimating the RERI

The RERI is defined as:

$$RERI = RR_{11} - RR_{10} - RR_{01} + 1, \quad (1)$$

where RR_{ab} is the relative risk (RR) in the group with X_1 exposure status a (1=exposed; 0=unexposed) and X_2 exposure status b (1=exposed; 0=unexposed) as compared to the doubly unexposed group. If we denote p_{ab} to be the probability of the outcome among the group of

subjects with X_1 equal to a and X_2 equal to b, then (1) can equivalently be written as the absolute risk due to interaction divided by the baseline risk (the risk in the doubly exposed group):

$$RERI = \frac{p_{11}}{p_{00}} - \frac{p_{10}}{p_{00}} - \frac{p_{01}}{p_{00}} + \frac{p_{00}}{p_{00}} = \frac{p_{11} - p_{10} - p_{01} + p_{00}}{p_{00}}$$

An RERI value of 0 implies no additive interaction, whereas values greater than 0 imply super-additive (positive) interaction and values less than 0 imply sub-additive (negative) interaction.

Although the RERI is defined in terms of relative risks (RRs), much of the literature evaluating the RERI uses odds ratios (ORs) from logistic regression models to approximate the relative risks (Hosmer et al., 1992; Assman et al., 1996; Vanderweele et al., 2012). This approximation is appropriate in studies where the outcome is rare, as is often true in case-control studies, or where incident cases are selected from a fixed cohort, controls are selected at the beginning of follow-up and censoring is unrelated to exposure (Knol et al., 2008). However, the OR overestimates the RR in other cases, and even slight overestimation of each RR can result in severe overestimation of the RERI (Zou et al., 2008). Thus, in many settings, it is important that RRs are used in estimating the RERI for assessing additive interaction.

A number of methods for deriving a confidence interval (CI) for the RERI have also been proposed, including the delta method (Hosmer et al., 1992), bootstrapping (Assman et al., 1996), and the method of variance estimates recovery (MOVER) (Zou et al., 2008). In simulations, Assman et al. (1996) found the symmetric delta method CIs were often completely below the true value in the scenarios with strong positive additive interaction, due to the right skewness of the RERI in this setting. The MOVER method is much less computationally intensive than the bootstrap procedure, and performed almost as well as the bootstrap in simulations (Zou et al., 2008). All of these approaches were studied in the independent data setting, and generally with very rare outcomes (e.g. $p_{00} = 0.00002$ in Assman et al.). As the prevalence of the outcome

increases, the RERI parameter space becomes more constrained (see A.1 in Appendix), which limits the extent of asymmetry in the sampling distribution. As a result, the delta method may provide appropriate coverage rates as long as outcomes are not extremely rare.

2.2.1 Extensions to population-averaged models

One approach for accounting for clustering in estimating the RERI is to utilize population-averaging models, in which the dependence among repeated measurements within clusters is considered a nuisance parameter. Accounting for this dependence structure can be accomplished via generalized estimating equations (GEEs) (Liang et al., 1986). Let K denote the number of clusters, n_k denote the number of observations for cluster k, k=1,..., K, and N denote the total sample size ($N = \sum_{k=1}^{K} n_k$). Let y_{ik} denote the binary outcome value for the i^{th} observation within the k^{th} cluster, and X_{1ik} and X_{2ik} denote the exposure status for two binary exposures of interest for the i^{th} observation within the k^{th} cluster (0=unexposed; 1=exposed). Lastly, let $\mathbf{C}_{ik}$ denote a vector of covariate values for the i^{th} observation within the k^{th} cluster. We assume the following form for the mean model:

$$\log\big(E(y_{ik})\big) = \beta_0 + \beta_1 X_{1ik} + \beta_2 X_{2ik} + \beta_3 X_{1ik} X_{2ik} + \boldsymbol{C}_{ik}\boldsymbol{\gamma} \qquad (2)$$

Under this model, the RERI is defined as $e^{\beta_1+\beta_2+\beta_3} - e^{\beta_1} - e^{\beta_2} + 1$. Previous research reported convergence problems for a log binomial model fit under a GEE framework (Pedroza et al., 2016). Alternatively, a modified Poisson approach can be used in clustered data settings, and provides reliable estimated RRs for studies with correlated binary data (Yelland et al. 2011; Zou et al., 2013). However, empirical coverage levels for CIs tend to be lower than the nominal level, particularly as the RRs and the within-cluster correlation increase. Thus, better

characterization of these models' performance in yielding appropriate estimates of the RERI is warranted.

2.2.2 Extensions to cluster-conditional models

In many clustered data settings, interest lies in characterizing variability across clusters or making cluster-specific predictions. Toward this aim, we fit a random intercept log binomial model, allowing the baseline probability p_{00} to vary by cluster:

$$\log\big(E(y_{ik})\big) = \beta_0^* + \beta_1^* X_{1ik} + \beta_2^* X_{2ik} + \beta_3^* X_{1ik} X_{2ik} + \boldsymbol{C}_{ik}^* \boldsymbol{\gamma} + b_{0k}, \qquad (3)$$

$$b_{0k} \sim N(0, \sigma_b^2)$$

where b_{0k} is the random deviation in intercept for cluster *k*. Assuming no unmeasured confounding conditional on cluster, the RERI is defined as in the population-averaged model: $e^{\beta_1^*+\beta_2^*+\beta_3^*} - e^{\beta_1^*} - e^{\beta_2^*} + 1$. Note that the RERI from the log binomial random intercepts model can be interpreted as a population-averaged RERI. That is, the cluster-conditional slope parameters are numerically equivalent to their respective marginal parameters under a log link, and therefore the cluster-conditional RERI is numerically equivalent to the marginal RERI (see A.2 in Appendix). This is an advantage of a log binomial random intercepts model over the logistic random intercepts model, even in the context of rare outcomes, since the cluster-conditional parameters are magnified relative to the marginal parameters under a logistic model (logit link) (Zeger et al., 1988; Neuhaus, 1992).

We are unaware of any literature exploring estimation of the RR for binary data under generalized linear mixed effects models to account for clustering in the frequentist setting. Torman and Camey successfully applied a Bayesian analysis of a log binomial random intercepts

model to a dataset for which the frequentist approach failed to converge, but did not investigate the operating characteristics of this approach under other settings (Torman et al., 2015).

In cluster-conditional models, including random slopes to allow the effects of particular covariates to vary by cluster may also be desirable. However, addition of such random slopes for the exposures would induce a distribution for the RERI measure itself; the RERI would vary by cluster and follow an unidentified distribution (the difference between two log normal distributions). This extension is beyond the scope of this paper.

2.3 Simulation Study

We performed a simulation study to investigate (1) what standard software packages could be used to reliably estimate the RERI from population-averaged and cluster-conditional regression models; (2) the bias of the estimated RERI as well as coverage and width of two different CI estimates for the RERI under the population-averaged log binomial and Poisson approximation models; and (3) the bias of the estimated RERI and the estimated standard deviation (SD) of the random intercept, as well as validity of inference on the RERI, across various implementations of the cluster-conditional model.

We assessed the performance of the different approaches across a range of baseline outcome prevalences. Table 2.1 defines the exposure/outcome scenarios. For each exposure/outcome scenario, 2,000 datasets were generated for 20, 50, and 275 clusters. Cluster sizes were generated from uniform distributions on (80, 200), (40, 80), and (1, 20), respectively, to give an average total sample size of 2,800-3,000. Additional simulations were performed on 275 clusters with cluster sizes generated from uniform distributions on (5, 20) and (30,50) to assess the effect of increasing cluster size while holding number of clusters constant.

Table 2.1 Exposure/Outcome Scenarios for the Simulation Study.

	Pr(outcome) by Exposure to X1, X2				Relative Risks (RRs)					
Scenario	Doubly-unexposed p_{00}	X1-exposed p_{10}	X2-exposed p_{01}	Doubly-exposed p_{11}	RR_{10}	RR_{01}	RR_{11}	Relative Excess Risk due to Interaction (RERI)	Relative Excess Odds due to Interaction (REOI)[a]	Random intercept standard deviation σ_b
1	0.01	0.02	0.03	0.09	2	3	9	5	5.71	0.61
2		0.02	0.03	0.14	2	3	14	10	12.03	0.5
3	0.1	0.1	0.1	0.2	1	1	2	1	1.25	0.4
4		0.2	0.3	0.5	2	3	5	1	3.89	0.17
5		0.1	0.2	0.5	1	2	5	3	6.75	0.17
6	0.2	0.2	0.2	0.4	1	1	2	1	1.67	0.23
7		0.2	0.4	0.6	1	2	3	1	3.33	0.13
8		0.4	0.4	0.8	2	2	4	1	11.67	0.05
9	0.4	0.4	0.4	0.8	1	1	2	1	5.00	0.05
10		0.4	0.4	0.6	1	1	1.5	0.5	1.25	0.13
11	0.6	0.6	0.6	0.9	1	1	1.5	0.5	5.00	0.02

Abbreviations: RR_{01}, relative risk of outcome in the group unexposed to X_1 and exposed to X_2 as compared to the doubly unexposed group; RR_{10}, relative risk of outcome in the group exposed to X_1 and unexposed to X_2 as compared to the doubly unexposed group; RR_{11}, relative risk of outcome in the doubly exposed group as compared to the doubly unexposed group; σ_b, standard deviation in random intercepts

[a] $REOI = OR_{11} - OR_{01} - OR_{10} + 1$, where OR_{ab} is the odds ratio comparing the group with exposure X_1=a and X_2=b to the doubly unexposed group.

For each parameter combination, we generated the i^{th} outcome from the k^{th} cluster from $y_{ik}|X_{1ik} = x_{ik}, X_{2ik} = x_{2ik} \sim Bernoulli\big(\pi(x_{1ik}, x_{2ik})\big)$ with the event probability $\pi(x_{1ik}, x_{2ik}) = \exp(\beta_0 + \beta_1 x_{1ik} + \beta_2 x_{2ik} + \beta_3 x_{1ik} x_{2ik} + b_{0k})$, where b_{0k} was generated under a Normal distribution with mean 0 and SD based on the scenario (see Table 2.1). Both X_1 and X_2 vary within cluster, and were assigned such that the proportion of being in the doubly exposed group, exposed only to X_1, and exposed only to X_2 were 0.10, 0.20, and 0.10, respectively.

2.3.1 Software implementations

To promote reporting of additive interaction effects for binary data in clustered data settings, we aimed to identify easy-to-use procedures and functions within familiar software programs. The population-averaged models were fit using a log binomial or modified Poisson model with an exchangeable covariance structure, implemented using the GENMOD procedure in SAS 9.4 (SAS Institute Inc., Carey, NC, USA). Final simulations for evaluating cluster-conditional models focused on a log binomial random intercept model and a Poisson random intercept model, both fit using the GLIMMIX procedure in SAS. Both pseudo-likelihood and Laplace approximation estimation techniques were considered (SAS/STAT User's Guide). Preliminary simulations also considered extending the COPY method (Deddens et al., 2003) and the McLaurin series approximation for estimation of the RR (Fitzmaurice et al., 2014) to cluster-conditional models with interaction, but convergence was no better than that of the frequentist standard log binomial random intercept model in preliminary simulations and they were not considered further.

Given the poor performance of the frequentist approaches to estimating the RERI in a mixed effect model including a random intercept, we also considered Bayesian methods. In particular, we jointly used the "brms"(Bürkner, 2017) and "rstan"(Stan Development Team) packages in R to fit a Bayesian log binomial random intercepts model in Stan using an R interface (see A.3 in Appendix). Two different weakly informative prior distributions were placed on the SD for the random intercepts: a half-Cauchy(0,5) and a Gamma(2,0.1) (Gelman, 2006; Chung, 2013).

Due to the increased computational resources required to fit the Bayesian models, they were fit on 500 simulated datasets for two specific scenarios. One scenario demonstrated poor convergence under frequentist methods (p_{00}=0.20, RR_{10}=2, RR_{01}=2 ,RR_{11}=4, RERI=1, σ_b=0.05) and one scenario demonstrated relatively good convergence under frequentist methods (p_{00}=0.20, RR_{10}=1, RR_{01}=1 ,RR_{11}=2, RERI=1, σ_b=0.23). We then fit the Bayesian model on the first 100 simulated datasets for all remaining scenarios. We ran four chains, each consisting of a 1,000 iteration burn-in period and a subsequent 1,000 iterations to estimate the posterior distribution. Simulation code in SAS and R is available online at: https://github.com/katcorr/Estimating-the-RERI-in-Clustered-Data-Settings.

2.3.2 Simulation results for population-averaged models

Both the log binomial and the Poisson GEE models converged for all scenarios with common outcomes (p_{00}≥0.10). The mean estimated RERIs and empirical coverage rates of 95% CIs for the RERI using the delta and MOVER methods for simulations with 275 clusters are shown in Table 2.2. Results from simulated datasets with 20 and 50 clusters are summarized in the Appendix (Table A2.1). The mean estimated RERIs were the same for the Poisson

approximation and the log binomial model, across cluster sizes and exposure/outcome scenarios. In general, they were unbiased, even with as few as 20 clusters. The exceptions were for rare outcomes and relatively large RERIs, where some upward bias was exhibited.

Table 2.2 The Mean Estimated Relative Excess Risk due to Interaction (RERI) and Empirical Coverage Rates of 95% Confidence Intervals for the RERI as Estimated From Generalized Estimating Equations.[a]

		Log Binomial GEE			Poisson GEE		
			95% CI Coverage			95% CI Coverage	
Baseline Outcome Prevalence	**RERI ($RR_{01}/RR_{10}/RR_{11}$)**	**Mean $\widehat{RERI}$**	**Delta**	**MOVER**	**Mean $\widehat{RERI}$**	**Delta**	**MOVER**
0.01	5 (2/3/9)	5.25	94.7	94.1	5.25	94.7	94.1
	10 (2/3/14)	10.60	95.4	93.7	10.60	95.4	93.8
0.1	1 (1/1/2)	1.00	95.1	95.5	1.00	95.1	95.5
	1 (2/3/5)	0.98	95.5	95.5	0.98	95.5	95.6
	3 (1/2/5)	3.00	95.5	95.4	3.00	95.5	95.4
0.2	1 (1/1/2)	1.00	95.5	95.5	1.00	95.5	95.5
	1 (1/2/3)	0.99	94.9	95.1	0.99	95.0	95.0
	1 (2/2/4)	1.00	94.9	95.0	1.00	95.0	95.1
0.4	1 (1/1/2)	1.00	95.2	95.1	1.00	95.1	95.2
	0.5 (1/1/1.5)	0.50	95.3	95.2	0.50	95.2	95.2
0.6	0.5 (1/1/1.5)	0.50	95.2	95.5	0.50	95.2	95.4

Abbreviations: CI, confidence interval; GEE, generalized estimating equations; MOVER, method of variance estimates recovery; RERI, relative excess risk due to interaction; RR_{01}, relative risk of outcome in the group unexposed to X_1 and exposed to X_2 as compared to the doubly unexposed group; RR_{10}, relative risk of outcome in the group exposed to X_1 and unexposed to X_2 as compared to the doubly unexposed group; RR_{11}, relative risk of outcome in the doubly exposed group as compared to the doubly unexposed group.

[a] Generalized estimating equations estimated specifying an exchangeable working correlation, log link, and either binomial or Poisson distribution, with robust variance estimates. Results are based on 2,000 simulated datasets, each with 275 clusters.

The empirical coverage rates of the 95% CIs for the RERI were very similar between the marginal log binomial and Poisson models, regardless of whether the delta method or MOVER method for CIs was employed. Coverage tended to be too low (in the 92-93% range) with 20 clusters, but was at the nominal level across all scenarios for simulations with 275 clusters. For baseline prevalences of 0.10 or more, the sampling distributions of the RERI were symmetric. Thus, the purported advantage of the MOVER method may not materialize when considering common outcomes. Furthermore, the MOVER CIs were slightly wider on average than the delta CIs for less common outcomes ($p_{00} < 0.40$) and had the same width for more common outcomes ($p_{00} \geq 0.40$) (data not shown).

Results from naive models (ignoring the clustering) suggested that more harm is done in using logistic regression than in ignoring the clustering, at least when there are two cluster-varying covariates. That is, among naive logistic models and marginal logistic models, there was substantial bias in the estimated RERIs and very poor nominal coverage rates for the 95% CIs. In contrast, the naïve log binomial model showed minimal bias in the estimated RERIs and 95% CIs contained the true values around the nominal level for the scenarios considered (Web Table 2). However, the naïve log binomial model had wider CIs than the marginal log binomial model (Web Table 3). Thus, when the exposures of interest vary within cluster, there is a gain in efficiency around the RERI estimate when accounting for the clustering.

2.3.3 Simulation results for cluster-conditional models

Convergence of the frequentist fit of the log binomial random intercept model was variable across cluster sizes and scenarios (Figure 2.1). Depending on the scenario, convergence occurred in between 49% and 99% of the simulations under 20 clusters, 33% and 98% of the

simulations under 50 clusters, and <1% to 74% under 275 clusters. The particularly poor convergence with 275 clusters (all but one scenario experienced <35% convergence) was presumed to be due to the small cluster sizes. Additional simulations conducted with increasing cluster size for simulations including 275 clusters (average cluster sizes of 21 and 40) showed improved convergence as cluster size increased, though the convergence under these larger sample sizes was still worse than with 20 and 50 clusters for some scenarios (data not shown). When using pseudo-likelihood estimation, the Poisson random intercept model occasionally failed to converge, whereas it converged consistently using Laplace estimation.

Among models that converged, both the frequentist log binomial and Poisson random intercept models underestimated the variability in the random intercepts. In many of the scenarios, the variability (SD) was estimated at zero, which reduces the model to a fixed-effects model and is not helpful toward (a) accounting for the within-cluster dependencies or (b) classifying the across-cluster variability (Figure 2.1). As the baseline outcome prevalence increased, the proportion of Poisson models that estimated the variability of the random intercepts to be greater than zero decreased, with 0% of the Poisson models being useable when the baseline outcome prevalence was 0.60. In fact, none of the log binomial models were useable when the baseline outcome prevalence was 0.60 for simulations with 275 clusters (and less than 15% were useable with 20 and 50 clusters).

Among the log binomial models that converged, the percent bias in the estimated RERI was generally negligible (mean bias <5%), except in the scenarios where there were very few useable models. For instance, the scenario with the largest absolute mean percent bias (-16%) was with 275 clusters, a baseline prevalence of 0.20, and an RERI of 3.0, where only 5% of the log binomial models converged and estimated σ_b^2 to be greater than 0.

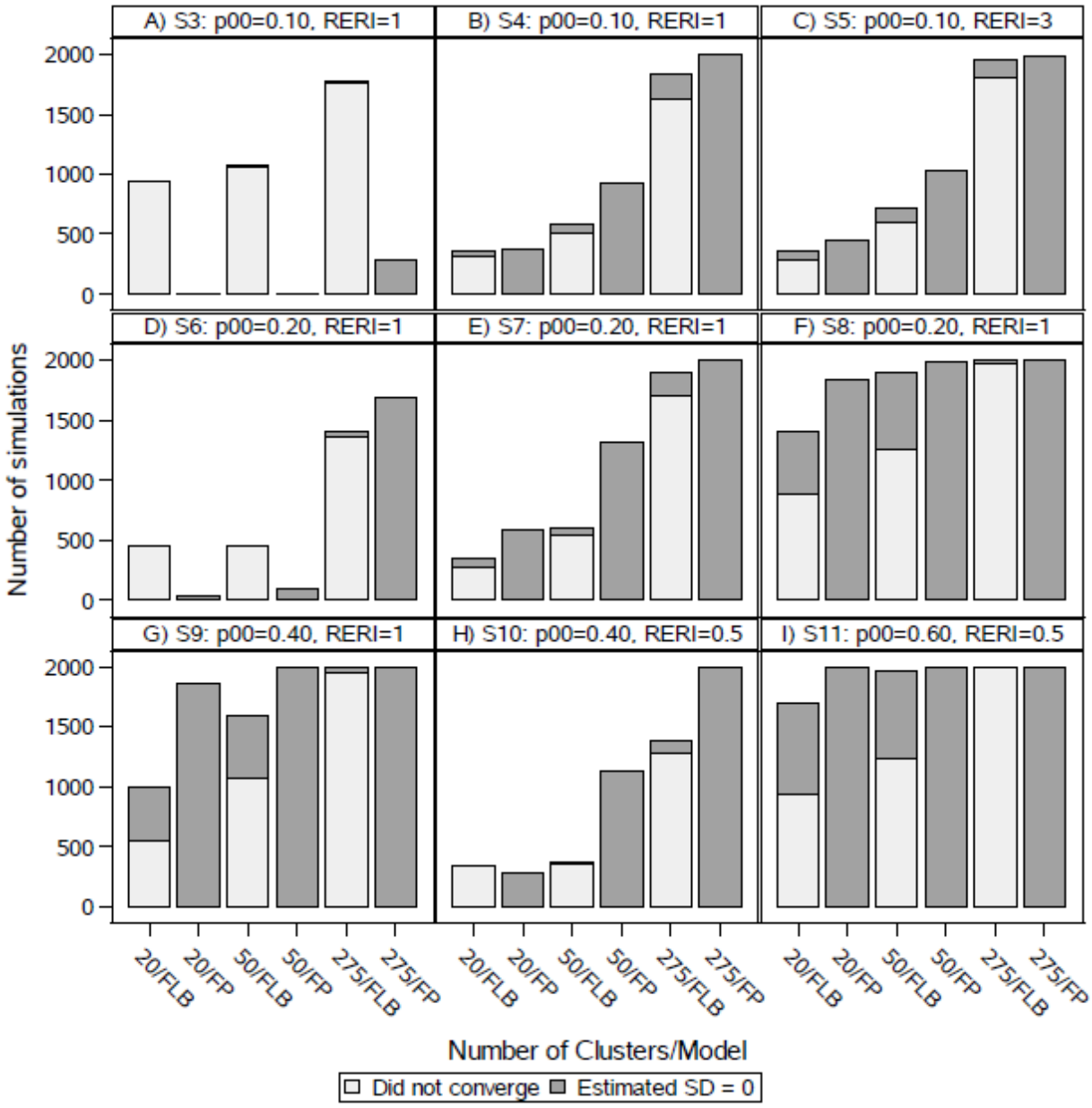

Figure 2.1 Convergence and degenerate estimates for the standard deviation of random intercepts in the frequentist log binomial (FLB) and frequentist Poisson (FP) random intercept models for Scenarios described in Table 2.1. Please refer to Table 2.1 to see the full scenario descriptions. Results are based on 2,000 simulated datasets per scenario. p00 is the outcome prevalence in the doubly unexposed group. For example, the fifth bar in panel D for scenario S7 is showing the results for the frequentist log binomial (FLB) random intercepts model for datasets simulated with 275 clusters (275/FLB): in 1,703 simulations, the model did not converge and, among the remaining models that did converge, 191 had a degenerate estimate for the variance of the random intercepts.

For simulations implementing the Bayesian fit of the log binomial random intercept model, the mean posterior RERIs and the mean posterior SDs in the random intercept were well calibrated (close to the true value) in most cases (Figures 2.2 and 2.3). For the scenario of a baseline prevalence of 0.20 and RRs of 1, 1 and 2 (RERI=1, Scenario (6)), the Bayesian fit always sampled and the mean posterior RERIs and SDs were estimated to be close to their true values. Moreover, for the scenario of a baseline prevalence of 0.20 and RRs of 2, 2, and 4 (RERI=1, Scenario (8)), the frequentist approach rarely yielded a model that converged and had a nondegenerate SD estimate. In contrast, the Bayesian posterior means for the RERI and SD were well-calibrated and the 95% credible intervals exhibited empirical coverage rates close to the nominal level. For instance, the true SD in the latter scenario was 0.05, and the mean of the posterior SD means ranged between 0.046 and 0.074 depending on the number of clusters and the prior distribution placed on the SD.

2.4 Application

Pregnant women with HIV are at higher risk for preterm delivery as compared to HIV-uninfected women, and exposure to certain antiretroviral therapies may increase the risk further (Watts et al., 2013). Globally, nevirapine is one of the most common therapies used in pregnant women, although it has been contraindicated in women with healthy immune function due to increased risk of hepatotoxicity (Fowler et al., 2016; U.S. Department of Health and Human Services). We consider data from the Surveillance Monitoring of ART Toxicities (SMARTT) study within the Pediatric HIV/AIDS Cohort Study (PHACS) network to evaluate potential

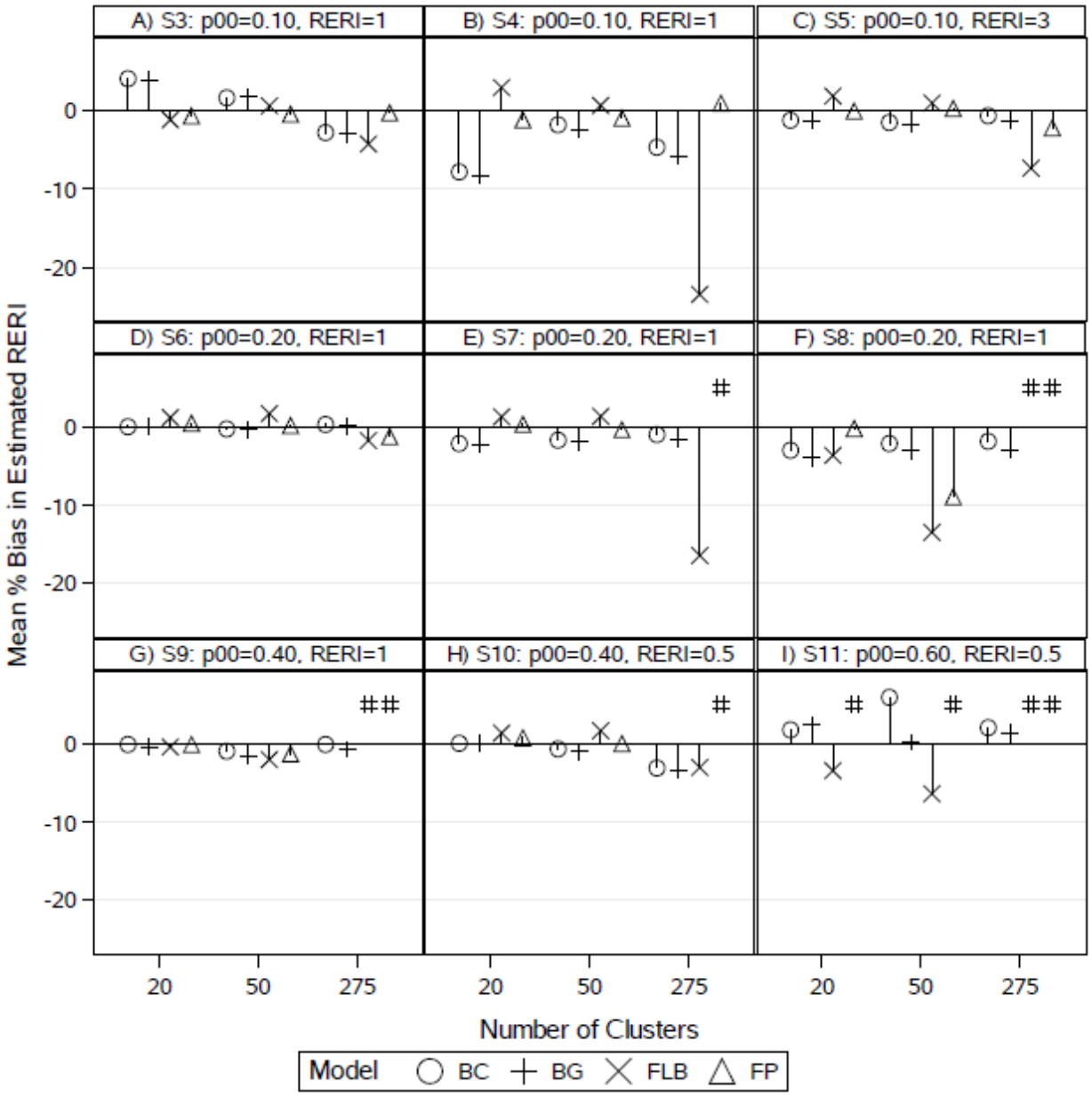

Figure 2.2 The mean % bias in estimated RERI across exposure/outcome scenarios and cluster sizes by model type for Scenarios described in Table 2.1. Please refer to Table 2.1 to see the full scenario descriptions. p00 is the outcome prevalence in the doubly unexposed group. BC indicates Bayesian log binomial random intercept with a half-Cauchy(0,5) prior distribution on the standard deviation (SD) for the random intercepts; BG indicates Bayesian log binomial random intercept with a Gamma(2,0.1) prior distribution on the SD; FLB indicates frequentist log binomial random intercept fit; and FP indicates frequentist Poisson random intercept model. # Note that some scenarios/clusters do not have markers for FLB/FP because there were no models that converged and had nondegenerate SD estimates under these fits.

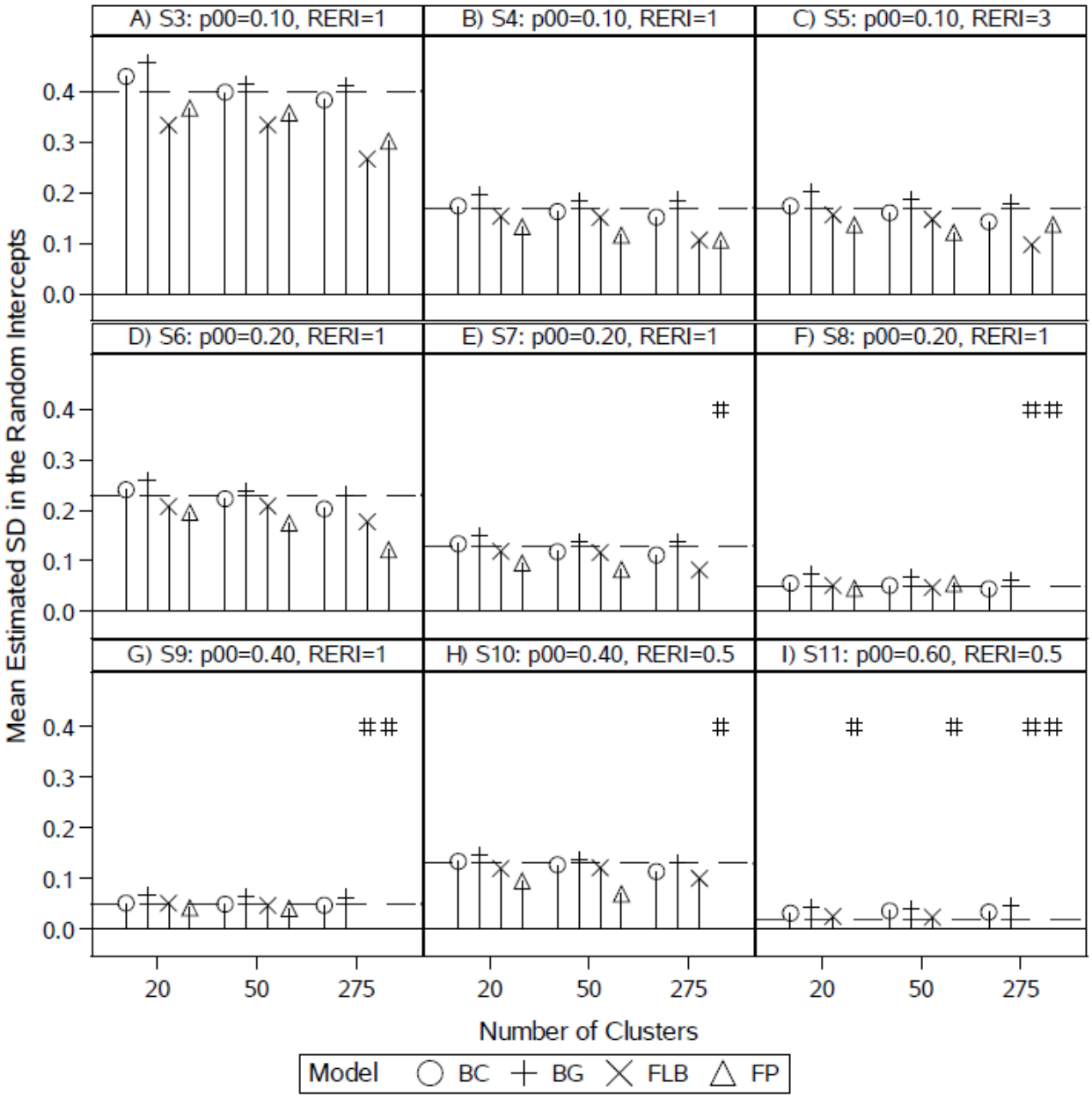

Figure 2.3 The mean estimated standard deviation (SD) in random intercepts across exposure/outcome scenarios and cluster sizes by model type for Scenarios described in Table 2.1. Please refer to Table 2.1 to see the full scenario descriptions. p00 is the outcome prevalence in the doubly unexposed group. BC indicates Bayesian log binomial random intercept with a half-Cauchy(0,5) prior distribution on the SD; BG indicates Bayesian log binomial random intercept with a Gamma(2,0.1) prior distribution on the SD; FLB indicates frequentist log binomial random intercept fit; and FP indicates frequentist Poisson random intercept fit. # Note that some scenarios/clusters do not have markers for FLB/FP because there were no models that converged and had nondegenerate SD estimates under these fits.

additive interaction between nevirapine exposure at conception and poor immunological health during pregnancy (earliest available CD4 count < 350 cells per cubic millimeter) on the risk of preterm delivery. The SMARTT cohort has enrolled over 3,000 HIV-infected pregnant women from 22 sites around the United States (Watts et al., 2013). A total of 3,202 women had information available on gestational age, antiretroviral therapy at conception, immunological health during pregnancy, and race, and were included in the analysis. The preterm delivery rates varied between 8.3% and 35.7% across sites (SD=6%), with between 14 and 335 women enrolled at a given site (average cluster size = 146).

Results from the unadjusted models suggested a strong positive additive interaction between nevirapine use at conception and poor immune function during pregnancy on preterm delivery (Table 2.3). In particular, the estimated RERI from the GEE log binomial model with a compound symmetry covariance structure was 1.78 (95% delta CI =0.69, 2.94); with a baseline risk of 17% in the doubly unexposed group, this amounts to an absolute risk due to interaction of 30% (17%*1.78). Observed results were consistent with expectations given the simulation results -- namely, (1) the CIs for the RERI from the naïve log binomial model (ignoring the clustering) were considerably wider than those from the GEE model (that is, precision was gained by accounting for the within-site correlations); (2) the MOVER CIs were wider than the delta method CIs across frequentist methods; (3) the CIs from the Poisson random intercept model were substantially wider than those from the log binomial random intercepts model, though the estimated SD for the random intercepts was only slightly lower; and (4) in the Bayesian analyses, the results were similar regardless of whether a half-Cauchy(0,5) or a Gamma(2,0.1) prior distribution was placed on the SD of the random intercepts, with the mean

of the estimated posterior distribution for the SD being slightly larger under the latter (Table 2.3).

Table 2.3 The Estimated Relative Excess Risk due to Interaction (RERI) between Nevirapine Use at Conception and Poor Immunological Health During Pregnancy on the Risk of Preterm Delivery Among 3,202 HIV-Infected Pregnant Women Across 22 Sites From the Surveillance Monitoring of ART Toxicities (SMARTT) Study Within the Pediatric HIV/AIDS Cohort Study (PHACS) Network, 1995-2015.

Model	$\widehat{RERI}$	$\hat{\sigma}_b$	**95% Confidence/Credible Intervals for RERI**	
Frequentist Marginal			**Delta CI**	**MOVER CI**
Naïve log binomial	1.87	N/A	0.44, 3.29	0.57, 3.57
GEE log binomial	1.78	N/A	0.58, 2.99	0.51, 3.06
Adjusted GEE log binomial[a]	1.82	N/A	0.83, 2.81	0.69, 2.77
Frequentist Conditional				
Log binomial random intercept	1.68	0.182	0.34, 3.01	0.43, 3.24
Poisson random intercept	1.75	0.165	-0.21, 3.70	0.13, 4.38
Adjusted Poisson random intercept[a]	1.45	0.163	-0.34, 3.23	-0.05, 3.85
Bayesian Conditional[b]			**Posterior Credible Interval**	
Log binomial random intercept (Cauchy[c])	1.49	0.193	0.19, 2.70	
Log binomial random intercept (Gamma[d])	1.46	0.213	0.23, 2.64	
Adjusted log binomial random intercept (Cauchy[c])	1.12	0.192	0.01, 2.23	
Adjusted log binomial random intercept (Gamma[d])	1.05	0.205	0.02, 2.10	

Abbreviations: $\hat{\sigma}_b$, estimated standard deviation in random intercepts; CI, confidence interval; GEE, generalized estimating equations with exchangeable correlation structure; MOVER, method of variance estimates recovery; N/A, not applicable

[a] Adjusted for black race and maternal age (<30, 30-39, 40+). The adjusted frequentist log binomial random intercept model did not converge.

[b] Results from unadjusted models are from four chains, each with 1,000 warmup samples and 1,000 post-warmup samples. Results from adjusted models are from four chains, each with 5,000 warmup samples and 5,000 post-warmup samples. $\widehat{RERI}$ is the mean posterior RERI and $\hat{\sigma}_b$ is the mean posterior standard deviation in the random intercepts.

[c] A half-Cauchy(0,5) prior distribution was placed on the standard deviation of the random intercepts.

[d] A Gamma(2,0.1) prior distribution was placed on the standard deviation of the random intercepts.

After adjusting for black race and maternal age at conception (<30, 30-39, 40+), the estimated RERI from the GEE model remained similar to that of the unadjusted model (RERI = 1.82 vs. 1.78). The adjusted frequentist log binomial random intercepts model did not converge. The adjusted Bayesian log binomial random intercepts model did not produce results when initial values were generated randomly using the default setting, but did sample when initial values were specified using informed estimates (see A.4 in Appendix). Additional iterations were required for chain convergence as assessed via potential scale reduction factors; 5,000 warm up and 5,000 post warm-up samples were used. Adjusted Bayesian results were attenuated compared to the unadjusted model (e.g. 1.46 versus 1.05 under a Gamma(2,0.01) prior on the SD).

2.5 Discussion

In examining the RERI metric for additive interaction in clustered data settings, it was important to consider that many dependent data settings assess outcomes with a much higher prevalence than the very rare outcomes assumed in previous RERI simulation studies. As such, it was important to estimate the RERI using RRs rather than ORs. While there has been much literature dedicated to estimating adjusted RRs for binary data, it has focused on the independent data setting and not in the context of estimating interaction effects (Wacholder, 1986; Deddens et al., 2003; McNutt et al., 2003; Barros et al., 2003; Carter et al., 2005; Spiegelman, et al., 2005; Yu et al., 2008; Chu et al., 2010; Marschner et al., 2012; Fitzmaurice et al., 2014; Lipsitz et al., 2015).

We found that estimating the RERI in log binomial or modified Poisson GEE models in clustered data settings was straightforward and efficient. In simulations and an application, there

were no problems with model convergence, and accounting for the within-cluster correlation increased precision around the RERI estimates as compared to a naïve model. We also found that the delta method provided valid CIs when the number of clusters was moderate to large. Given that the delta method requires less computation than the MOVER method and yields CIs that are generally slightly narrower yet provide similar coverage as the MOVER method, it appears that computing confidence limits for the RERI using the delta method is appropriate and advantageous in the clustered data setting for population-averaged models. When the number of clusters is small (20-50), the coverage rates are lower than desirable; bootstrapping may be advantageous in these settings if the computational resources and time required to bootstrap is not prohibitive for a given application.

In contrast to the marginal models, there were difficulties in fitting the frequentist log binomial models with random intercepts. The observed patterns suggest that convergence is affected by both number of clusters and cluster size. We found worse convergence with increasing number of clusters, which is opposite what we had expected. A possible explanation is that, with increased number of clusters, there is more opportunity for very large (or very small) random intercepts, which could push some observed probabilities outside the parameter space.

The Poisson approximation with robust standard errors has been a common approach to estimating RRs for binary data. However, we found that a Poisson random intercepts model often severely underestimated the SD in the random effects, frequently reducing it to a fixed effects model. Furthermore, it produced overly conservative confidence intervals for the RERI, even when using robust standard errors for the regression parameters.

Despite the additional computational resources required, the Bayesian approach to fitting the log binomial random intercepts model offers several advantages. In the simulation study,

there were no issues with sampling, the posterior mean RERI was well-calibrated (unbiased), the posterior mean SD for the random intercepts exhibited less bias than the frequentist approaches, and Bayesian inference was straightforward with valid credible intervals for the RERI. In the data application, the adjusted frequentist log binomial model did not converge, but with additional work – specifying informed initial values for the chains and increasing the number of iterations per chain -- the adjusted Bayesian log binomial random intercepts model sampled and yielded reasonable results.

Our simulations had not considered any confounding factors, and the instability of the log binomial random intercepts model is likely to increase with the addition of covariates. Furthermore, all models assumed that the random intercepts were normally distributed, as assumed in PROC GLIMMIX. Assuming a normal distribution for the random intercepts in a log binomial model ignores the parameter constraints on the log probabilities. A different prior distribution that recognizes that constraint may be more appropriate. It would be difficult to fit a frequentist model using existing software assuming a non-Normal distribution for the random intercepts, but the Bayesian approach is more amenable to such updates. More research is warranted on estimating the RERI from cluster-conditional models with random slopes and additional covariates.

In summary, when assessing interaction between exposures in clustered data settings, the RERI can be estimated from frequentist log binomial GEE models or Bayesian log binomial random intercept models, depending on additional aims of the analysis (e.g. estimation of cluster-to-cluster variability). Using the log linear as opposed to logit link model is particularly important for accurate estimation of the RERI, even when the background outcome prevalence is as low as 10%.

3. A computationally efficient algorithm for permutation testing of rare genetic variants

3.1 Introduction

With the rapid advancement in DNA sequencing technologies over the last decades, cost-effective identification of rare and very rare single-nucleotide polymorphisms (SNPs) has become technically feasible and reliable in large scale association studies. Yet, detecting associations between single rare variants and specific diseases remains a challenge (Lee et al., 2014; Auer et al, 2015; Zhang, 2015). Rare and very rare variants are typically defined as an observed minor allele frequency of <5% and <1%, respectively. Consequently, unless a study has a very large number of participants or there is a very strong association between a variant and disease, standard statistical tests will inherently have low statistical power to detect associations with single rare variants. Furthermore, asymptotic tests that rely on large sample theory may not be able to maintain the specified type I error.

Statistical methods have been proposed to assess associations between rare variants and disease which address these issues by collapsing genotypes across variants or grouping rare variants by location (Morgenthaler 2007; Li 2008; Wu 2011). However, such methods could adversely impact power if a single underlying disease susceptibility locus is grouped together with null-loci. Moreover, grouping variants does not allow for identification of specific variant-disease associations.

For single rare variants, Ma et al. (2011) recommended Firth's likelihood ratio test for low count variants, while acknowledging that the test is not well calibrated in studies with extremely rare variants (expected minor allele count < 40; minor allele frequency (MAF) < 0.001) (Ma

2011; Firth 1993). Recently, Dey et al. (2017) proposed a score-based test that uses the saddlepoint approximation to estimate the null distribution of the score statistic when it is far from the mean. They showed that their test, which allows for covariate adjustment, is much faster than Firth's test and controls the type I error even with extremely unbalanced case-control ratios. Although better than the normal approximation, the test still relies upon an approximation. Furthermore, the test was developed for phenome-wide association studies which often have extremely unbalanced case-control ratios, and it was not developed particularly for rare variant testing.

In this paper we introduce a computationally efficient algorithm for permutation testing between a single rare variant and affection status, which also allows inclusion of covariates (e.g. principal components to adjust for population substructure and epidemiological variables such as age, sex, etc.). A special feature of the proposed algorithm is that the implementation of the random permutation of the genotype vector has only a numerical complexity of O(N*p), where the parameter N is the sample size and p is the allele frequency. This, in combination with a very efficient computation of the score functions, enables permutation testing at a genome-wide level with the required significance level of 10^{-8} and smaller. To illustrate the feasibility of the approach, we apply the method to a study of chronic obstructive pulmonary disease (COPD). In simulations, we show that the permutation test maintains a type I error rate closer to the nominal level than the asymptotic and saddlepoint approximation tests.

3.2 Materials and Methods

We consider the setting of a case-control study in which *N* subjects are sequenced on *m* rare genetic variants (MAF < 0.05). For each genetic variant, we are interested in testing H_0: $\beta_j = 0$ versus H_A: $\beta_j \neq 0$ (j = 1, 2, … m) in the following model:

$$logit(p_{ij}) = \boldsymbol{c_i^T \alpha_j} + \beta_j G_{ij},$$

where p_{ij} is the probability of being a case (P($y_i = 1$)) for the i^{th} individual, conditional on the allele count for the j^{th} genetic variant and the covariates in $\boldsymbol{c}$; G_{ij} is the number of minor alleles for the i^{th} individual and the j^{th} genetic variant; and $\boldsymbol{c_i}$ is a p x 1 column vector of covariate values for the i^{th} individual, including an intercept term. Note that, although the proportion of cases may not reflect the population prevalence of disease, logistic regression is suitable to test a genetic variant in a case-control study (Prentice et al., 1979). The score function with respect to β_j is:

$$U_j = \sum_{i=1}^{N} G_{ij}(y_i - \tilde{p}_{ij})$$

The variance of the score under H_0 is:

$$Var(U_j) = \boldsymbol{g_j^T W g_j} - \boldsymbol{g_j^T W C (C^T W C)^{-1} C^T W g_j},$$

where $\boldsymbol{g_j}$ is the N *x 1* vector of minor allele counts for the j^{th} genetic variant, $\boldsymbol{W}$ is an N *x* N diagonal matrix with $Var(y_i)|_{H_0} = \tilde{p}_i(1 - \tilde{p}_i)$ on the i^{th} diagonal element, $\tilde{p}_i$ is the probability of being a case for the i^{th} individual in the null model (conditional on only the covariates in $\boldsymbol{c}$), and $\boldsymbol{C} = [\boldsymbol{c_1}\ \boldsymbol{c_2} \ldots \boldsymbol{c_n}]^{\boldsymbol{T}}$.

We can test H_0: $\beta_j = 0$ by evaluating the score statistic $S_j = \frac{U_j^2}{Var(U_j)}$ under the null hypothesis. Asymptotically, S_j follows a chi-square distribution with one degree of freedom. However, with rare variants, large sample theory may not hold and p-values relying on the assumption that $S_j \sim \chi_1^2$ may be invalid. Instead, we propose testing H_0: $\beta_j = 0$ via a computationally efficient permutation test of the score function for β_j. That is, the distribution of U_j^2 under the null is empirically estimated by permuting G_{ij}, calculating the permuted score

$U_{j,PRM}^2$ for each permutation, and comparing the observed value of U_j^2 (based on the observed data) to the empirical null distribution. Alternatively, one could use standardized score tests, i.e. $\frac{U_j^2}{Var(U_j)}$ (see Appendix A.6). The standardized score could potentially be more powerful when the true null distribution and the asymptotic distribution become similar, but it comes at a cost of substantially increased numerical complexity during the permutation tests. We investigated the power of U_j^2 and the standard score test. Our results suggest that there is no advantage of using the standardized score over U_j^2 for rare variant data. Given the substantially higher numerical burden of the standardized score, we recommend the use of U_j^2 and implemented the proposed algorithm accordingly.

The number of rare variants is often quite large (>10^6), and the number of permutations per variant will need to be large (generally on the order of 10^9 to investigate the small levels around genome-wide significance) in order to reliably estimate the null distribution and obtain valid p-values. An advantage of using the score statistic to test β_j is that the score test is evaluated under the constrained model (where $\beta_j = 0$) and thus does not involve fitting a logistic regression model with $\boldsymbol{g_j}$. The estimated constrained risk is independent of the genetic variant being tested. Furthermore, we only permute the genetic variant in the permutations so that the relationship between disease and confounding variables in the observed dataset is preserved in the replicate datasets. The $\tilde{p}_i$s used in calculating the permuted score statistics are thus also the same as the $\tilde{p}_i$s used in computing the observed score statistic. That is, only one logistic regression model needs to be fit, regardless of the number of variants being tested and the number of permutations being conducted.

Nevertheless, repeating a permutation test requiring 10^9 permutations across a large number of variants will be computationally intensive and could become prohibitive for the large numbers of rare variants often assessed in genetic studies today. Hecker et al. (2018, manuscript in preparation) proposed a sequential testing approach to permutation-based association testing that directly tests the permutation-based p-value against a pre-specified significance level. This approach can drastically reduce the number of permutations required for the majority of variants; for instance, using a significance level of 5×10^{-8}, they found that the 99.92% of least significant variants needed an average of 22 permutations per variant. We developed an algorithm that is efficient and robust in terms of the numerical and computational aspects, e.g. drawing *m* out of *N* without replacement. In conjunction with the sequential testing approach, this algorithm enables permutation testing between a single rare variant and affection status, adjusting for covariates, to be implemented on a whole-genome wide scale.

3.2.1. The algorithm

The fact that we are analyzing rare variants implies that $\boldsymbol{g_j}$ is a sparse vector. We use this to our advantage in calculating the permuted scores. In particular, note that:

$$U_j = \sum_{i=1}^{N} G_{ij}(y_i - p_{ij}) = \sum_{i \in M_{1j}} (y_i - p_i) + 2 * \sum_{i \in M_{2j}} (y_i - p_i),$$

where M_{1j} is the set of n_{1j} subjects with one minor allele on the j^{th} genetic variant and M_{2j} is the set of n_{2j} subjects with two minor alleles on the j^{th} genetic variant. The *j* index is dropped from p_{ij} since the constrained disease risks do not depend on the genetic variant as noted above.

Our algorithm proceeds as follows:

1. Fit the constrained logistic regression model, and compute the residual vector $\tilde{\boldsymbol{r}} = (\boldsymbol{y} - \tilde{\boldsymbol{p}})$.
2. For each variant (j=1,2, …, m):
    a. Generate a vector $\boldsymbol{q_j}$ of length $n_{1j} + n_{2j}$, with n_{1j} ones and n_{2j} twos
    b. Compute the square of the observed score.
        i. Subset $\tilde{\boldsymbol{r}}$ on the M_{1j} and M_{2j} indices; call this subset vector $\tilde{\boldsymbol{r}}^{(j0)}$
        ii. Compute $U_j^2 = [\boldsymbol{q_j^T}\tilde{\boldsymbol{r}}^{(j0)}]^{\mathbf{2}}$
    c. Compute the square of the permuted scores. For k in 1 to K (where K is the number of permutations):
        i. Randomly select without replacement $n_{1j} + n_{2j}$ values between 1 and N using a modified Durstenfeld shuffle (Durstenfeld, 1964; see Appendix A.5)
        ii. Subset $\tilde{\boldsymbol{r}}$ on the randomly selected values; call this subset vector $\tilde{\boldsymbol{r}}^{(jk)}$
        iii. Calculate the square of the permuted score for the k^{th} permutation for the j^{th} variant: $[U_{j,PRM}^k]^{\mathbf{2}} = [\boldsymbol{q_j^T}\tilde{\boldsymbol{r}}^{(jk)}]^{\mathbf{2}}$
    d. Calculate the p-value as the proportion of permuted scores that are as extreme or more extreme than the observed score: $p - value_j = \frac{\sum_{k=1}^{K} I([U_{j,PRM}^k]^2 \geq U_j^2)}{(K+1)}$

Analyses were performed in R 3.2.1 (R Core Team). The saddlepoint approximation test was conducted using the SPAtest package (Dey et al., 2017) `ScoreTest_SPA` function. Code is available at: https://github.com/katcorr/Permutation-Test-for-Rare-Genetic-Variants.

3.3 Results

3.3.1 COPD Study: Feasibility

To demonstrate the feasibility of our permutation algorithm, we used the whole genome-sequencing data available on 1,794 participants enrolled in two COPD studies, the COPDgene study (Regan et al., 2010) and the Boston Early-Onset COPD (BEOCOPD) study (Silverman et al, 2013). The COPDgene study has enrolled over 10,000 smokers with and without COPD between 45 and 80 years old. The BEOCOPD study has enrolled over 200 severe, early-onset COPD patients (less than 53 years old with forced expiratory volume in one second (FEV1) < 40%) and their family members. As part of a Trans-Omics for Precision Medicine (TOPMed) program sponsored by the National Heart, Lung and Blood Institute, cases were selected from the COPDgene study and the BEOCOPD study and controls were selected from the COPDgene study for sequencing. We analyzed 631,244 genetic variants on the 22^{nd} chromosome for 821 cases and 973 controls.

Almost half of the variants had only one minor allele (MAF=0.00028). An additional 9% of the variants had only two minor alleles (MAF=0.00056), and 80% of the variants had a MAF < 1%.

We fit the constrained logistic regression model adjusting for 10 principal components based on the Jaccard index to identify population stratification (Prokopenko et al., 2018, manuscript in preparation). Since our sample only included 1,794 participants, we were able to compute the exact null distribution of the score statistic for variants with one or two non-zero minor alleles. We computed the null distributions using the same computations as in the permutation algorithm, except instead of generating random values for each permutation, we ran the algorithm through every possible genetic variant vector. For the remaining SNPs, we ran the

algorithm as described in the Methods section. In addition to the permutation test, we calculated p-values from the standard asymptotic test (assuming a χ_1^2 distribution for the score statistic) and the fastSPA-2 test (Dey et al., 2017).

The distributions of p-values as generated by the fastSPA-2 test versus the permutation test are compared in Figure 3.1 for variants with one non-zero minor allele count. For permuted p-values < 0.05, the fastSPA-2 p-values were more conservative than necessary. Note that the fastSPA-2 p-value and the asymptotic p-value were equivalent for all variants with MAC=1, as the observed score statistic did not fall outside two standard deviations of the mean.

Although we cannot expect to identify extremely rare variants at a genome-wide significance level given our sample size, we could discover a suggestion of a global contribution of very rare variants to COPD if the observed score distribution for the rare variants shows an unusual number of large score statistics relative to the null distribution. We did not find such a suggestion between COPD and variants on chromosome 22 with MACs of one or two alleles (Figure 3.2).

For other variants with MAF < 1%, the fastSPA-2 p-values were not consistently above or below the respective permutation p-values. For variants with MAF between 1 and 5% and for common variants (MAF > 5%), the asymptotic, fastSPA-2 and permuted p-values were similar.

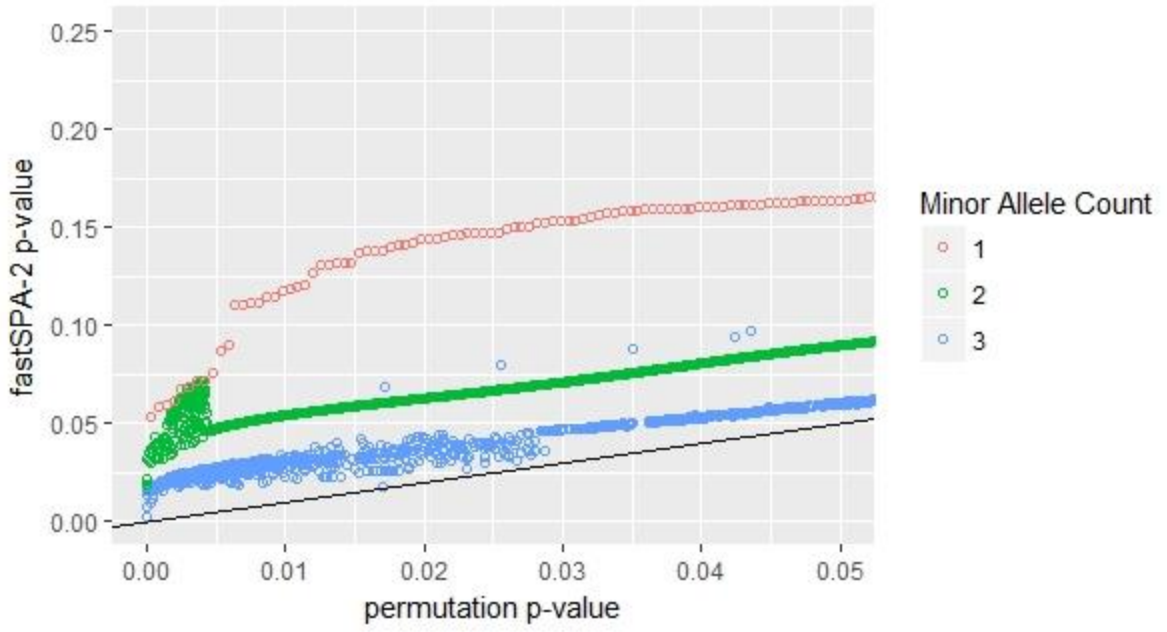

Figure 3.1 P-values from the fastSPA-2 test versus the permutation test for variants on chromosome 22 with minor allele counts of one to three with permuted p-values < 0.05.

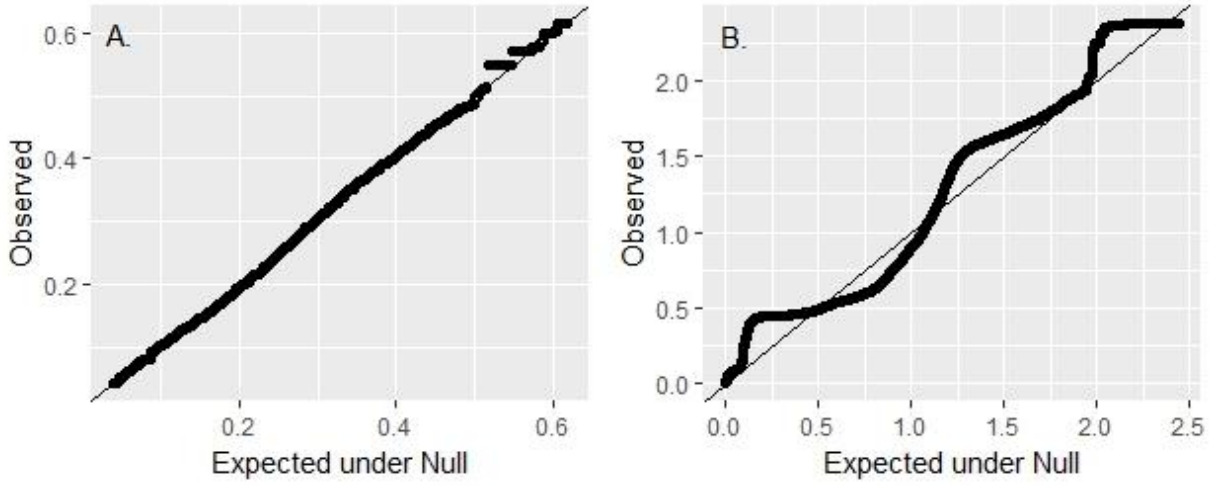

Figure 3.2 QQ-plots comparing the distribution of the score statistic under the null hypothesis and the observed score statistics for variants with one (A) or two (B) minor allele counts.

3.3.2 Simulation study

We conducted a simulation study to demonstrate the feasibility of our permutation algorithm and to identify scenarios where the approximate tests fail to control Type I error. We generated 100,000 replicate datasets with sample size 20,000 and overall disease prevalence of 10%. We varied the case-control ratio (1:1, 1:2, 1:4, 1:16, 1:100, and 1:400) and the minor allele frequency (between 0.00005 and 0.01). We also generated datasets under the alternative hypothesis (OR=1.05, 1.5 and 2.0) at various sample sizes to assess power. For each permutation test, 10^5 permutations were used.

The type I error rates at the 5% level across scenarios and tests are shown in Table 3.1. Under extremely unbalanced case-control ratios, the type I error rate for the asymptotic and fastSPA-2 tests were inflated. For instance, with a MAF of 1% and a case-control ratio of 1:400, the type I error rates for the asymptotic and fastSPA-2 tests were 0.0601 and 0.0590, respectively. In this scenario, the permutation test remained valid, although conservative (type I error rate=0.0183 at 5% level). With a MAF of 0.1% and a case-control ratio of 1:400, the type I error rates for the asymptotic and fastSPA-2 tests were severely inflated at 0.0964 and 0.0962, respectively. The permutation test again remained valid and conservative (type I error rate=0.0050 at 5% level).

For extremely rare variants (MAF $\leq$ 0.05%) and reasonably balanced case-control ratios, the fastSPA-2 and permutation tests maintained the type I error rate better than the asymptotic test. For example, under a MAF of 0.005% and a 1:4 case-control ratio, the type I error rate for the asymptotic test was > 0.10, but < 0.05 for the fastSPA-2 and permutation tests. With a MAF of 0.05% and a 1:1 case-control ratio, the type I error rate was 0.0520 for both the asymptotic

and fastSPA-2 tests, 0.0297 for the conservative permutation test, and 0.0457 for the permutation test that weighted equality of permuted and observed scores at one-half.

Table 3.1 A comparison of observed type I error rates for the asymptotic test, fastSPA-2 test, and permutation test, across 100,000 simulated datasets of sample size 20,000.

MAF	Case-Control Ratio	Type I Error, α=0.05		
		Asymptotic	fastSPA-2	Permuted
0.01	1:1	0.0488	0.0488	0.0434
	1:2	0.0492	0.0492	0.0492
	1:4	0.0489	0.0489	0.0492
	1:16	0.0490	0.0491	0.0500
	1:100	0.0540	0.0533	0.0510
	1:400	0.0601	0.0589	0.0182
0.005	1:1	0.0490	0.0490	0.0411
	1:2	0.0504	0.0504	0.0516
	1:4	0.0503	0.0503	0.0519
0.001	1:1	0.0495	0.0495	0.0350
	1:2	0.0497	0.0497	0.0511
	1:4	0.0486	0.0486	0.0517
	1:16	0.0399	0.0399	0.0329
	1:100	0.0596	0.0596	0.0175
	1:400	0.0964	0.0961	0.0050
0.0005	1:1	0.0520	0.0520	0.0297
	1:2	0.0490	0.0490	0.0443
	1:4	0.0458	0.0458	0.0492
0.0001	1:1	0.0423	0.0072	0.0053
	1:2	0.0394	0.0239	0.0223
	1:4	0.0615	0.0412	0.0194
0.00005	1:1	0.0164	0.0006	0.0006
	1:2	0.0456	0.0108	0.0107
	1:4	0.1029	0.0276	0.0177

Abbreviations: fastSPA-2, fast saddlepoint approximation test; MAF, minor allele frequency

A comparison of the power of each test at a sample size of 5,000 and true OR of 1.5 is shown in Table 3.2. The power of the permutation test based on U_j alone is the same as the power of the permutation test based on the standardized score, S_j. With a 1:2 case-control ratio, the power of the permutation test is higher than that of the fastSPA-2 test. In particular, at a MAF of 0.005, the power is about 42% under the permutation test and only 34% under the fastSPA-2 test.

Table 3.2 A comparison of power for the asymptotic test, fastSPA-2 test, and permutation tests, across 100,000 simulated datasets of sample size 5,000 with OR=1.5.

MAF	Case-Control Ratio	Power					
		Asymptotic	fastSPA-2	Permutation Test			
				U_j[a]	U_j[b]	S_j[c]	S_j[d]
0.01	1:1	0.5868	0.5868	0.5505	0.5841	0.5505	0.5841
	1:2	0.6475	0.5591	0.6493	0.6493	0.6493	0.6493
	1:4	0.4559	0.4559	0.4833	0.4833	0.4833	0.4833
0.005	1:1	0.3354	0.3354	0.2909	0.3310	0.2909	0.3310
	1:2	0.3949	0.3360	0.4162	0.4165	0.4162	0.4165
	1:4	0.2726	0.2726	0.2743	0.2752	0.2743	0.2752

Abbreviations: fastSPA-2, fast saddlepoint approximation test; MAF, minor allele frequency

[a] Based on U_j and calculated as $\frac{\sum_{k=1}^{K} I([U_{j,PRM}^k]^2 \geq U_j^2)}{(K+1)}$

[b] Based on U_j and calculated as $\frac{\sum_{k=1}^{K} I([U_{j,PRM}^k]^2 > U_j^2) + 0.5 * \sum_{k=1}^{K} I([U_{j,PRM}^k]^2 = U_j^2)}{(K+1)}$

[c] Based on S_j and calculated as $\frac{\sum_{k=1}^{K} I(S_{j,PRM}^k \geq S_j)}{(K+1)}$

[d] Based on S_j and calculated as $\frac{\sum_{k=1}^{K} I(S_{j,PRM}^k > S_j) + 0.5 * \sum_{k=1}^{K} I(S_{j,PRM}^k = S_j)}{(K+1)}$

3.4 Discussion

For rare genetic variants, asymptotic tests that depend on large sample theory can have inflated type I error rates. Estimating the distribution of a test statistic using permutations is a suitable alternative, but permutation tests can be computationally expensive and the amount of time and resources required can become prohibitive. We developed a computationally efficient algorithm to perform permutation testing of rare genetic variants that allows for adjustment of covariates. The algorithm takes advantage of the sparsity of exposure; with decreasing MAF, the computation time of the permutation algorithm decreases. Thus, the algorithm is fastest where needed most -- in cases where the asymptotic and approximate tests are most questionable (extremely rare variants).

In a COPD sequencing study and in simulations, we showed the feasibility of the permutation testing algorithm. Although we cannot expect to identify extremely rare variants at a genome-wide significance level under reasonable sample sizes, the permutation test could allow the discovery of a global contribution of very rare variants on an outcome if the observed score distribution for the rare variants shows an unusual number of large score statistics relative to the permuted sampling distribution.

In the simulation study we observed scenarios where both the asymptotic and fastSPA-2 tests break down (e.g. MAF of 0.1% and 1:400 case-control ratio observed type I error rates of >9% at the 5% level). In contrast, the permutation test maintained the type I error level at or below 0.05, even under extremely rare MAF. Due to the discrete nature of the null distribution of the score statistic in the case of extremely rare variants, the empirical type I error levels for the permutation test were often well below 0.05.

Additional work toward further speeding up the permutation algorithm is of interest, including implementing the algorithm in C/C++ and integrating a sequential testing strategy into the permutation algorithm (Hecker et al., 2018, manuscript in preparation) to discard clearly non-significant variants early.

Conclusions

There are many avenues of further research that spawn from this dissertation research. In Chapter 1, we observed the importance of adjusting for other spontaneous drug exposures in screening the safety of antiretroviral therapies taken during pregnancy by women with HIV infection. The hierarchical modeling approach as studied was implemented using a frequentist approach. After having observed the advantages a Bayesian approach can offer in mixed effect regression settings with common, binary outcomes in Chapter 2, it may be worthwhile to evaluate the Chapter 1 model (without interactions) in a Bayesian framework and with a log (rather than logit) link. Whereas the frequentist implementation is limited by software options (e.g. must assume a Normal distribution for the random effects in PROC GLIMMIX), a Bayesian approach can be more flexible.

Another logical next step is to extend the hierarchical model from Chapter 1 to screen for additive drug-drug or drug-covariate interactions -- as originally intended -- under a Bayesian framework. Another bridging of the research in Chapters 1 and 2 could involve incorporating clustered outcomes into a hierarchical model that groups drugs by drug class. That is, in the motivating application presented in Chapter 1, we accounted for other ARV exposures in screening the safety of individual ARVs taken during pregnancy on the risk of preterm delivery, but we did not account for the fact that patients were clustered by clinic. In the example application presented in Chapter 2, we accounted for the clustering of patients within clinics – and assessed the variability in preterm delivery across clinics – but our model only included the ARV of interest (nevirapine) and did not account for other ARV exposures. Further research could focus on methods which both adjust for other ARV exposures and account for correlated

outcomes. Correlated outcomes could arise from patients clustered within clinic and/or multiple pregnancies from the same woman.

Lastly, there is still room to reduce the computational time and resources required to implement the permutation testing algorithm presented in Chapter 3 on a genome-wide scale. In addition to coding the algorithm in C/C++, which is much faster than R – or using an R interface to C++ (e.g., Rcpp) -- incorporating a sequential testing strategy into the permutation algorithm to discard clearly nonsignificant variants early would reduce the computational burden.

Appendices

A.1 The parameter space of the RERI

Conditional on the baseline probability p_{00}, the RERI ($RERI = \frac{p_{11}}{p_{00}} - \frac{p_{10}}{p_{00}} - \frac{p_{01}}{p_{00}} + 1$) is bounded. Assuming exposures are coded such that the baseline group represents the group with the lowest risk (Knol et al., 2011), the lower and upper bounds of the RERI can be derived as:

$$2 - \frac{2}{p_{00}} \leq RERI \leq \frac{1}{p_{00}} - 1.$$

The minimum possible RERI is attained when RR_{11} is at its minimum and RR_{10} and RR_{01} are at their maximums. Assuming as noted above, that the baseline group represents the group with the lowest risk ($p_{00} \leq p_{01}, p_{10}, p_{11}$), the minimum value of RR_{11} is 1.0 and occurs when $p_{11} = p_{00}$. The maximum values of RR_{10} and RR_{01} are attained when p_{01} and p_{10} are at their maximum value (equal to 1). Therefore, a lower bound for the RERI is: $1 - \frac{1}{p_{00}} - \frac{1}{p_{00}} + 1 = 2 - \frac{2}{p_{00}}$.

Similarly, the maximum possible RERI is attained when RR_{11} is at its maximum and RR_{10} and RR_{01} are at their minimums. Assuming again that $p_{00} \leq p_{01}, p_{10}, p_{11}$, the maximum value of RR_{11}occurs when $p_{11} = 1$. The minimum values for RR_{10} and RR_{01} occur when $p_{10} = p_{01} = p_{00}$. Thus, an upper bound for the RERI is: $\frac{1}{p_{00}} - 1 - 1 + 1 = \frac{1}{p_{00}} - 1$.

For example, if $p_{00} = 0.20$, then $-8 \leq RERI \leq 4$; and if $p_{00} = 0.40$, then $-3 \leq RERI \leq 1.5$. For rarer outcomes, the possible parameter space of the RERI becomes much wider (eg., if $p_{00} = 0.01$, then $-198 \leq RERI \leq 99$).

A.2 Numerical equivalence of cluster-conditional RERI and induced marginal RERI under log binomial random intercepts model

The cluster-conditional log binomial mean model is:

$$\log\big(E(y_{ik}|b_{0k})\big) = \beta_0^* + \beta_1^* X_{1ik} + \beta_2^* X_{2ik} + \beta_3^* X_{1ik} X_{2ik} + \boldsymbol{C}_{ik}^* \boldsymbol{\gamma} + b_{0k},$$

$$b_{0k} \sim N(0, \sigma_b^2)$$

and the cluster-conditional RERI is thus defined as:

$$RERI_{CC} = \exp(\beta_1^* + \beta_2^* + \beta_3^*) - \exp(\beta_1^*) - \exp(\beta_2^*) + 1$$

The induced marginal mean model is:

$$E_Y(y_{ik}) = E_b\big(E_Y(y_{ik}|b_{0k})\big)$$

$$= E_b(\exp(\beta_0^* + \beta_1^* X_{1ik} + \beta_2^* X_{2ik} + \beta_3^* X_{1ik} X_{2ik} + \boldsymbol{C}_{ik}^* \boldsymbol{\gamma} + b_{0k}))$$

$$= E_b(\exp(\beta_0^* + \beta_1^* X_{1ik} + \beta_2^* X_{2ik} + \beta_3^* X_{1ik} X_{2ik} + \boldsymbol{C}_{ik}^* \boldsymbol{\gamma}) \exp(b_{0k}))$$

$$= \exp(\beta_0^* + \beta_1^* X_{1ik} + \beta_2^* X_{2ik} + \beta_3^* X_{1ik} X_{2ik} + \boldsymbol{C}_{ik}^* \boldsymbol{\gamma}) * E_b(\exp(b_{0k}))$$

$$= \exp(\beta_0^* + \beta_1^* X_{1ik} + \beta_2^* X_{2ik} + \beta_3^* X_{1ik} X_{2ik} + \boldsymbol{C}_{ik}^* \boldsymbol{\gamma}) * \exp(\sigma_b^2/2)$$

$$= \exp((\beta_0^* + \sigma_b^2/2) + \beta_1^* X_{1ik} + \beta_2^* X_{2ik} + \beta_3^* X_{1ik} X_{2ik} + \boldsymbol{C}_{ik}^* \boldsymbol{\gamma}),$$

where the second to last line follows from the fact that $b_{0k} \sim N(0, \sigma_b^2)$ so $\exp(b_{0k})$ is log-normally distributed with mean $\exp(\sigma_b^2/2)$. And, thus, the marginal RERI induced by the cluster-conditional model is:

$$RERI_M = \exp(\beta_1^* + \beta_2^* + \beta_3^*) - \exp(\beta_1^*) - \exp(\beta_2^*) + 1$$

As the slopes in the cluster-conditional model are the same after averaging over the clusters, the induced marginal slopes are numerically equivalent to the cluster-conditional slopes and the cluster-conditional RERI is therefore numerically equivalent to the marginal RERI.

This numerical equivalence is specific to the random intercepts model with log link (e.g. log binomial or Poisson approximation). If normally-distributed random slopes are included in the model, or if a different link function is used (e.g. logit link for a logistic random intercepts model), the induced marginal RERI from the cluster-conditional model may not be the same as the cluster-conditional RERI. For instance, consider a model with random intercepts and slopes:

$$\log\big(E(y_{ik}|\boldsymbol{b})\big) = \beta_0^* + \beta_1^* X_{1ik} + \beta_2^* X_{2ik} + \beta_3^* X_{1ik}X_{2ik} + \boldsymbol{C}_{ik}^*\boldsymbol{\gamma} + b_{0k} + b_{1k}X_{1ik} + b_{2k}X_{2ik},$$

$$\boldsymbol{b_k} \sim N\left(\boldsymbol{0}, \begin{bmatrix} \sigma_{b0}^2 & 0 & 0 \\ 0 & \sigma_{b1}^2 & 0 \\ 0 & 0 & \sigma_{b2}^2 \end{bmatrix}\right)$$

In this case, the induced marginal mean model is:

$$E_Y(y_{ik}) = E_{\boldsymbol{b}}\big(E_Y(y_{ik}|\boldsymbol{b_k})\big)$$

$$= E_{\boldsymbol{b}}(\exp(\beta_0^* + \beta_1^* X_{1ik} + \beta_2^* X_{2ik} + \beta_3^* X_{1ik}X_{2ik} + \boldsymbol{C}_{ik}^*\boldsymbol{\gamma} + b_{0k} + b_{1k}X_{1ik} + b_{2k}X_{2ik}))$$

$$= E_{\boldsymbol{b}}(\exp(\beta_0^* + \beta_1^* X_{1ik} + \beta_2^* X_{2ik} + \beta_3^* X_{1ik}X_{2ik} + \boldsymbol{C}_{ik}^*\boldsymbol{\gamma})\exp(b_{0k})\exp(b_{1k}X_{1ik})\exp(b_{2k}X_{2ik}))$$

$$= \exp(\beta_0^* + \beta_1^* X_{1ik} + \beta_2^* X_{2ik} + \beta_3^* X_{1ik}X_{2ik} + \boldsymbol{C}_{ik}^*\boldsymbol{\gamma}) * E_{\boldsymbol{b}}(\exp(b_{0k})) * E_{\boldsymbol{b}}(\exp(b_{1k}X_{1ik}))$$

$$* E_{\boldsymbol{b}}(\exp(b_{2k}X_{2ik}))$$

$$= \exp(\beta_0^* + \beta_1^* X_{1ik} + \beta_2^* X_{2ik} + \beta_3^* X_{1ik}X_{2ik} + \boldsymbol{C}_{ik}^*\boldsymbol{\gamma}) * \exp(\sigma_{b0}^2/2) * [\exp\left(\frac{\sigma_{b1}^2}{2}X_{1ik}^2\right)]$$

$$* [\exp\left(\frac{\sigma_{b2}^2}{2}X_{2ik}^2\right)]$$

$$= \exp\left((\beta_0^* + \sigma_{b0}^2/2) + \beta_1^* X_{1ik} + \beta_2^* X_{2ik} + \beta_3^* X_{1ik}X_{2ik} + \boldsymbol{C}_{ik}^*\boldsymbol{\gamma} + \frac{\sigma_{b1}^2}{2}X_{1ik}^2 + \frac{\sigma_{b2}^2}{2}X_{2ik}^2\right)$$

$$= \exp\left((\beta_0^* + \sigma_{b0}^2/2) + (\beta_1^* + \frac{\sigma_{b1}^2}{2}X_{1ik})X_{1ik} + (\beta_2^* + \frac{\sigma_{b2}^2}{2}X_{2ik})X_{2ik} + \beta_3^* X_{1ik}X_{2ik} + \boldsymbol{C}_{ik}^*\boldsymbol{\gamma}\right)$$

Clearly, the slopes for the exposures of interest (X_{1ik}, X_{2ik}) in the cluster-conditional model are not the same as the respective slopes in the induced marginal model, and thus the two RERIs will not be equivalent either.

A.3 Implementation of a Bayesian log binomial random intercepts model

We jointly used the "brms" and "rstan" packages in R to fit a Bayesian log binomial random intercepts model. The "brms" package provides a user-friendly interface to fit Bayesian generalized linear mixed models using Stan. The formula syntax is similar to that of R's popular "lme4" package, and the sampling scheme is extremely efficient and thus fast in terms of Bayesian computation. A log link is not allowed to be specified for a binary outcome in the "brms" package. However, the "make_stanmodel" and "make_standata" functions within the "brms" package were used to help specify the Stan file and data. In particular, the "make_stanmodel" function was used to output Stan model code for a random intercepts logistic model. The resulting Stan code was updated to change the model from a logit link to a log link and to calculate the RERI directly. The "make_standata" function was used to create a list of the Stan data for the respective Stan model. Then, the "stan" function within the "rstan" package was called to run the Stan model on the Stan data.

R Code

```
library("brms")
library("rstan")
rstan_options(auto_write=TRUE)
options(mc.cores=parallel::detectCores())

# make_stancode and make_standata functions are from BRMS package
stan.logit <- make_stancode(outcome ~ x1 + x2 + x1:x2 + (1|Cluster)
                    , data=simdata
                    , family="bernoulli"(link="logit")
                    , iter=1000
                    # set gamma(2,0.1) prior on SD
                    , prior = set_prior("gamma(2,0.1)", class = "sd")
                      )
standata <- make_standata(outcome ~ x1 + x2 + x1:x2 + (1|Cluster)
                                  , data=simdata)
```

```
# edit Stan code output from make_stancode
# so using log link instead of logit and save file as
# "bayes_logbin_gamma.stan" (included below)

stfit <- stan(file="bayes_logbin_gamma.stan"
              , data=standata, chains=4, iter=2000)
```

Stan Code for log binomial random intercepts model with a Gamma(2,0.1) prior on SD

```
// generated with brms 1.6.1
functions {
}
data {
  int<lower=1> N;  // total number of observations
  int Y[N];  // response variable
  int<lower=1> K;  // number of population-level effects
  matrix[N, K] X;  // population-level design matrix
  // data for group-level effects of ID 1
  int<lower=1> J_1[N];
  int<lower=1> N_1;
  int<lower=1> M_1;
  vector[N] Z_1_1;
  int prior_only;  // should the likelihood be ignored?
}
transformed data {
  int Kc;
  matrix[N, K - 1] Xc;  // centered version of X
  vector[K - 1] means_X;  // column means of X before centering
  Kc = K - 1;  // the intercept is removed from the design matrix
  for (i in 2:K) {
    means_X[i - 1] = mean(X[, i]);
    Xc[, i - 1] = X[, i] - means_X[i - 1];
  }
}
parameters {
  vector[Kc] b;  // population-level effects
  real temp_Intercept;  // temporary intercept
  vector<lower=0>[M_1] sd_1;  // group-level standard deviations
  vector[N_1] z_1[M_1];  // unscaled group-level effects
}
transformed parameters {
  // group-level effects
  vector[N_1] r_1_1;
  r_1_1 = sd_1[1] * (z_1[1]);
}
model {
  vector[N] mu;
  mu = Xc * b + temp_Intercept;
  for (n in 1:N) {
    mu[n] = mu[n] + (r_1_1[J_1[n]]) * Z_1_1[n];
  }
  // prior specifications
  sd_1 ~ gamma(2,0.1);
  z_1[1] ~ normal(0, 1);
  // likelihood contribution
  if (!prior_only) {
```

```
    Y ~ bernoulli(exp(mu));
  }
}
generated quantities {
  real b_Intercept;  // population-level intercept
  real RERI;         // relative excess risk due to interaction
  b_Intercept = temp_Intercept - dot_product(means_X, b);
  RERI = exp(b[1]+b[2]+b[3]) - exp(b[1]) - exp(b[2]) + 1;
}
```

A.4 Specification of initial values for the adjusted Bayesian log binomial random intercepts model

Initial estimates for the beta coefficients were generated from a multivariate normal distribution with means equal to the respective estimated coefficients from the frequentist adjusted modified Poisson GEE model and covariance matrix equal to the covariance matrix of the coefficients. The modified Poisson GEE model was used to allow more variability in initial estimates across chains (as compared to a log binomial GEE model).

The initial values for the standard deviation in the random intercept were generated from a uniform distribution ranging between the 2.5^{th} and 97.5^{th} percentiles for the standard deviation as estimated from the unadjusted Bayesian log binomial random intercepts model.

A.5 Randomly select without replacement $n_{1j} + n_{2j}$values between 1 and N

Since $\boldsymbol{g_j}$ is a sparse vector (rare variant), instead of shuffling the entire vector, we randomly select without replacement $n_{1j} + n_{2j}$ values between 1 and N using a modified version of Durstenfeld's shuffle (Durstenfeld, 1964) which shuffles in place and has time complexity O($n_{1j} + n_{2j}$). Let $m = n_{1j} + n_{2j}$.

```
sample.custom <- function(N,m){
    vector <- c(1:N)                                        O(1)
    for (i in 0:(m-1)){                                     O(m)
        num <- round(runif(n=1,min=0,max=1)*(N-i))              O(1)
        vector[c(N-i,num)] <- vector[c(num,N-i)]                O(1)
    }
    return(vector[(N-m+1):N])                               O(1)
}
```

$$O(1) + [O(m) * O(1)] + O(1) = O(1) + O(m) + O(1)$$

$$= O(m),$$

where the first equality follows from the fact that O(f)*O(g) = O(f*g), and the second equality follows from the fact that O(f) + O(g) = O(max(f,g)) and $m \geq 1$.

A.6 Permutation algorithm based on S_j

The calculations for the variance of U_j can be reduced by noting that:

$$Var(U_j) = \boldsymbol{g_j^T W g_j} - \boldsymbol{g_j^T W C (C^T W C)^{-1} C^T W g_j}$$

$$= \boldsymbol{g_j^T W g_j} - \boldsymbol{g_j^T A g_j},$$

where $\boldsymbol{A} = \boldsymbol{WC(C^TWC)^{-1}C^TW}$, and both $\boldsymbol{W}$ and $\boldsymbol{A}$ are constant across variants and across permutations. Since $\boldsymbol{W}$ is a diagonal matrix, the first term in $Var(U_j)$ can be reduced in a similar manner as U_j above:

$$\boldsymbol{g_j^T W g_j} = \sum_{i=1}^{N} G_{ij}^2 p_{ij}(1-p_{ij}) = \sum_{i \in M_{1j}} p_i(1-p_i) + 4 * \sum_{i \in M_{2j}} p_i(1-p_i)$$

We can further take advantage of the sparsity of $\boldsymbol{g_j}$ and note that:

$$\boldsymbol{g_j^T A g_j} = \boldsymbol{g_j^{*T} A^* g_j^*},$$

where $\boldsymbol{g_j^*}$ is an $(n_{1j} + n_{2j})$ *x 1* vector of the non-zero minor allele counts for the j^{th} genetic variant and $\boldsymbol{A}^*$ is an $(n_{1j} + n_{2j})x(n_{1j} + n_{2j})$ matrix keeping the rows and columns of $\boldsymbol{A}$ corresponding to the non-zero minor allele counts for the j^{th} genetic variant as indexed by M_{1j} and M_{2j}.

The permutation algorithm based on S_j proceeds similar to the permutation algorithm based on U_j^2, with the additional computations for the variance included:

1. Fit the constrained logistic regression model, and compute the residual vector $\tilde{\boldsymbol{r}} = (\boldsymbol{y} - \tilde{\boldsymbol{p}})$, the variance vector $\tilde{\boldsymbol{v}} = \tilde{\boldsymbol{p}}(\boldsymbol{1} - \tilde{\boldsymbol{p}})$, $\widetilde{\boldsymbol{W}}$=diag($\tilde{\boldsymbol{v}}$), and $\boldsymbol{A} = \widetilde{\boldsymbol{W}}\boldsymbol{C}(\boldsymbol{C^T}\widetilde{\boldsymbol{W}}\boldsymbol{C})^{-1}\boldsymbol{C^T}\widetilde{\boldsymbol{W}}$.
2. For each variant (j=1,2, …, m):
    a. Generate a vector $\boldsymbol{q_j}$ of length $n_{1j} + n_{2j}$, with n_{1j} ones and n_{2j} twos

b. Calculate $\boldsymbol{q_j^2}$, with n_{1j} twos and n_{2j} fours (element-wise squaring of $\boldsymbol{q_j}$)

c. Compute the observed score statistic.

   i. Subset $\tilde{\boldsymbol{r}}$ on the M_{1j} and M_{2j} indices; call this subset vector $\tilde{\boldsymbol{r}}^{(j0)}$

   ii. Subset $\tilde{\boldsymbol{v}}$ on the M_{1j} and M_{2j} indices; call this subset vector $\tilde{\boldsymbol{v}}^{(j0)}$

   iii. Create $\boldsymbol{A^{(j0)}}$ by keeping only the rows and columns of $\boldsymbol{A}$ equal to the M_{1j} and M_{2j} indices

   iv. Compute $U_j = \boldsymbol{q_j^T \tilde{r}^{(j0)}}$ and $Var(U_j) = \boldsymbol{(q_j^2)^T \tilde{v}^{(j0)} - q_j^T A^{(j0)} q_j}$

   v. Compute $S_j = \frac{\boldsymbol{(q_j^T \tilde{r}^{(j0)})^2}}{\boldsymbol{(q_j^2)^T \tilde{v}^{(j0)} - q_j^T A^{(j0)} q_j}}$

d. Compute the permuted score statistics. For k in 1 to K (where K is the number of permutations):

   i. Randomly select without replacement $n_{1j} + n_{2j}$values between 1 and N

   ii. Subset $\tilde{\boldsymbol{r}}$ on the randomly selected values; call this subset vector $\tilde{\boldsymbol{r}}^{(jk)}$

   iii. Subset $\tilde{\boldsymbol{v}}$ on the randomly selected values; call this subset vector $\tilde{\boldsymbol{v}}^{(jk)}$

   iv. Create $\boldsymbol{A^{(jk)}}$ by keeping only the rows and columns of $\boldsymbol{A}$ equal to the randomly selected values

   v. Calculate the permuted score for the k^{th} permutation for the j^{th} variant:

   $$S_{j,PRM}^{k} = \frac{\boldsymbol{(q_j^T \tilde{r}^{(jk)})^2}}{\boldsymbol{(q_j^2)^T \tilde{v}^{(jk)} - q_j^T A^{(jk)} q_j}}$$

Calculate the p-value as the proportion of permuted scores that are as extreme of more extreme than the observed score: $p - value_j = \frac{\sum_{k=1}^{K} I(S_{j,PRM}^{k} \geq S_j)}{(K+1)}$

Table A1.1 Scenario (i): bias, efficiency, and coverage of 95% confidence intervals across statistical approaches for the binary outcome under scenario (i) (no true effects). Results are from 3,000 simulations each with sample size 1,000.

		Mean Bias in Log Odds			Mean SE of Log Odds			Coverage of 95% CI		
% Exposed	Drug	Separate[a]	Full[b]	Hierarchical	Separate[a]	Full[b]	Hierarchical	Separate[a]	Full[b]	Hierarchical
<5%	EFV	0.185	0.213	-0.012	0.870	0.924	0.517	0.96	0.96	0.99
	ETR	-0.001	0.010	-0.016	0.759	0.795	0.489	0.96	0.96	0.98
	NVP	-0.047	-0.049	-0.009	0.668	0.727	0.477	0.96	0.97	0.98
	FPV	0.059	0.084	0.019	0.794	0.851	0.450	0.96	0.97	0.99
	NFV	-0.068	-0.052	0.021	0.489	0.592	0.408	0.97	0.96	0.98
5-15%	ABC	-0.026	-0.051	0.003	0.303	0.440	0.306	0.95	0.95	0.97
	RPV	-0.069	-0.056	-0.012	0.437	0.520	0.424	0.96	0.96	0.96
	DRV	-0.031	-0.014	0.012	0.286	0.388	0.341	0.95	0.95	0.96
>15%	3TC	0.001	0.100	0.020	0.196	0.721	0.345	0.95	0.96	1.00
	FTC	-0.004	-0.078	0.019	0.196	0.947	0.351	0.95	0.97	0.99
	TDF	-0.004	0.167	0.016	0.196	0.942	0.347	0.94	0.96	0.99
	ZDV	0.000	-0.001	0.016	0.198	0.483	0.305	0.95	0.95	0.98
	ATV	-0.007	0.015	0.023	0.223	0.356	0.322	0.96	0.96	0.96
	LPV/r	-0.012	0.001	0.020	0.217	0.385	0.328	0.95	0.95	0.96

Abbreviations: ABC, Abacavir; ATV, Atazanavir; CI, confidence interval; DRV, Darunavir; EFV, Efavirenz; ETR, Etravirine; FPV, Fosamprenavir; FTC, Emtricitabine; LPV/r, Ritonavir-boosted Lopinavir; NFV, Nelfinavir; NVP, Nevirapine; RPV, Rilpivirine; SE, standard error; TDF, Tenofovir; ZDV, Zidovudine; 3TC, Lamivudine.

[a] 23.3% of the EFV models, 10.2% of the ETR models, 14.0% of the FPV models, 0.2% of the NFV models, 4.9% of the NVP models, and 0.1% of the RPV models did not converge or yielded unreasonable SEs. Results presented exclude these models.

[b] 22.0% of the full models did not converge. Results presented exclude these models.

Table A1.2 Scenario (ii): bias, efficiency, and coverage of 95% confidence intervals across statistical approaches for the binary outcome under scenario (ii) (all protease inhibitors have a subtle effect on preterm birth). Results are from 3,000 simulations each with sample size 1,000.

		Mean Bias in Log Odds			Mean SE of Log Odds			Coverage of 95% CI		
% Exposed	Drug	Separate[a]	Full[b]	Hierarchical	Separate[a]	Full[b]	Hierarchical	Separate[a]	Full[b]	Hierarchical
<5%	EFV	-0.076	0.211	-0.017	0.870	0.919	0.511	0.98	0.96	0.99
	ETR	-0.121	-0.029	-0.022	0.723	0.761	0.476	0.98	0.97	0.98
	NVP	-0.325	-0.080	-0.022	0.671	0.725	0.471	0.99	0.97	0.98
	FPV	-0.280	-0.020	0.017	0.733	0.791	0.437	0.99	0.97	0.99
	NFV	-0.289	-0.025	0.021	0.424	0.528	0.392	0.95	0.96	0.98
5-15%	ABC	-0.139	-0.031	0.008	0.283	0.413	0.294	0.94	0.95	0.98
	RPV	-0.321	-0.053	-0.017	0.431	0.510	0.416	0.94	0.96	0.96
	DRV	-0.257	-0.003	0.017	0.253	0.363	0.325	0.86	0.95	0.97
>15%	3TC	0.029	0.093	0.028	0.178	0.663	0.334	0.95	0.95	1.00
	FTC	-0.023	-0.009	0.024	0.178	0.893	0.339	0.95	0.97	0.99
	TDF	-0.023	0.094	0.024	0.178	0.898	0.336	0.95	0.97	0.99
	ZDV	0.012	0.000	0.020	0.179	0.452	0.293	0.95	0.95	0.98
	ATV	-0.240	0.009	0.018	0.200	0.339	0.310	0.79	0.95	0.96
	LPV/r	-0.232	0.007	0.021	0.194	0.362	0.317	0.79	0.94	0.96

Abbreviations: ABC, Abacavir; ATV, Atazanavir; CI, confidence interval; DRV, Darunavir; EFV, Efavirenz; ETR, Etravirine; FPV, Fosamprenavir; FTC, Emtricitabine; LPV/r, Ritonavir-boosted Lopinavir; NFV, Nelfinavir; NVP, Nevirapine; RPV, Rilpivirine; SE, standard error; TDF, Tenofovir; ZDV, Zidovudine; 3TC, Lamivudine.

[a] 22.5% of the EFV models, 7.3% of the ETR models, 6.8% of the FPV models, 0.1% of the NFV models, 4.5% of the NVP models, and 0.1% of the RPV models did not converge or yielded unreasonable SEs. Results presented exclude these models.

[b] 15.1% of the full models did not converge. Results presented exclude these models.

Table A1.3 Scenario (iii.a): bias, efficiency, and coverage of 95% confidence intervals across statistical approaches for the binary outcome under scenario (iii.a) (LPV/r has a moderate effect on preterm birth). Results are from 3,000 simulations each with sample size 1,000.

		Mean Bias in Log Odds			Mean SE of Log Odds			Coverage of 95% CI		
% Exposed	Drug	Separate[a]	Full[b]	Hierarchical	Separate[a]	Full[b]	Hierarchical	Separate[a]	Full[b]	Hierarchical
<5%	EFV	-0.045	0.208	-0.004	0.870	0.923	0.513	0.99	0.97	0.99
	ETR	-0.186	0.012	-0.011	0.749	0.786	0.481	0.98	0.97	0.98
	NVP	-0.275	-0.064	-0.008	0.667	0.722	0.471	0.98	0.97	0.98
	FPV	-0.160	0.049	0.169	0.791	0.847	0.442	0.98	0.96	0.98
	NFV	-0.332	-0.067	0.102	0.493	0.588	0.400	0.97	0.96	0.98
5-15%	ABC	-0.184	-0.021	-0.030	0.291	0.416	0.295	0.94	0.95	0.98
	RPV	-0.283	-0.031	0.013	0.432	0.515	0.419	0.95	0.95	0.96
	DRV	-0.285	-0.009	0.078	0.282	0.386	0.335	0.86	0.95	0.96
>15%	3TC	0.320	0.087	0.037	0.181	0.685	0.335	0.58	0.95	0.99
	FTC	-0.315	-0.011	-0.005	0.182	0.906	0.341	0.59	0.97	0.99
	TDF	-0.315	0.107	-0.004	0.181	0.916	0.338	0.59	0.96	0.99
	ZDV	0.322	0.004	0.039	0.180	0.462	0.294	0.57	0.95	0.98
	ATV	-0.305	0.010	0.077	0.218	0.354	0.318	0.74	0.96	0.96
	LPV/r	-0.001	0.027	-0.083	0.186	0.358	0.312	0.94	0.95	0.95

Abbreviations: ABC, Abacavir; ATV, Atazanavir; CI, confidence interval; DRV, Darunavir; EFV, Efavirenz; ETR, Etravirine; FPV, Fosamprenavir; FTC, Emtricitabine; LPV/r, Ritonavir-boosted Lopinavir; NFV, Nelfinavir; NVP, Nevirapine; RPV, Rilpivirine; SE, standard error; TDF, Tenofovir; ZDV, Zidovudine; 3TC, Lamivudine.

[a] 24.0% of the EFV models, 9.7% of the ETR models, 13.2% of the FPV models, 0.3% of the NFV models, 4.1% of the NVP models, and 0.0% of the RPV models did not converge or yielded unreasonable SEs. Results presented exclude these models.

[b] 17.2% of the full models did not converge. Results presented exclude these models.

Table A1.4 Scenario (iii.b): bias, efficiency, and coverage of 95% confidence intervals across statistical approaches for the binary outcome under scenario (iii.b) (ABC has a moderate effect on preterm birth). Results are from 3,000 simulations each with sample size 1,000.

		Mean Bias in Log Odds			Mean SE of Log Odds			Coverage of 95% CI		
% Exposed	Drug	Separate[a]	Full[b]	Hierarchical	Separate[a]	Full[b]	Hierarchical	Separate[a]	Full[b]	Hierarchical
<5%	EFV	0.101	0.238	-0.067	0.865	0.912	0.507	0.97	0.96	0.99
	ETR	-0.093	0.015	-0.064	0.755	0.790	0.482	0.97	0.96	0.98
	NVP	-0.022	-0.075	-0.061	0.638	0.694	0.461	0.96	0.97	0.98
	FPV	0.103	0.018	-0.049	0.758	0.821	0.437	0.95	0.96	0.99
	NFV	-0.142	-0.049	-0.076	0.482	0.574	0.395	0.98	0.96	0.98
5-15%	ABC	-0.017	-0.009	-0.202	0.247	0.387	0.286	0.96	0.95	0.91
	RPV	-0.139	-0.046	-0.068	0.430	0.507	0.411	0.97	0.96	0.97
	DRV	-0.037	-0.015	-0.046	0.276	0.374	0.327	0.95	0.95	0.96
>15%	3TC	0.184	0.088	0.210	0.189	0.673	0.343	0.84	0.96	0.97
	FTC	-0.180	0.034	0.099	0.189	0.913	0.350	0.85	0.97	0.99
	TDF	-0.167	0.056	0.104	0.189	0.884	0.342	0.86	0.96	0.98
	ZDV	0.057	0.006	0.061	0.190	0.432	0.294	0.93	0.95	0.97
	ATV	-0.062	0.006	-0.045	0.217	0.344	0.308	0.94	0.95	0.95
	LPV/r	-0.103	0.003	-0.089	0.213	0.366	0.313	0.93	0.95	0.95

Abbreviations: ABC, Abacavir; ATV, Atazanavir; CI, confidence interval; DRV, Darunavir; EFV, Efavirenz; ETR, Etravirine; FPV, Fosamprenavir; FTC, Emtricitabine; LPV/r, Ritonavir-boosted Lopinavir; NFV, Nelfinavir; NVP, Nevirapine; RPV, Rilpivirine; SE, standard error; TDF, Tenofovir; ZDV, Zidovudine; 3TC, Lamivudine.

[a] 23.7% of the EFV models, 11.5% of the ETR models, 10.7% of the FPV models, 0.4% of the NFV models, and 2.5% of the NVP models did not converge or yielded unreasonable SEs. Results presented exclude these models.

[b] 13.9% of the full models did not converge. Results presented exclude these models.

Table A1.5 Scenario (iii.c): bias, efficiency, and coverage of 95% confidence intervals across statistical approaches for the binary outcome under scenario (iii.c) (EFV has a moderate effect). Results are from 3,000 simulations each with sample size 1,000.

		Mean Bias in Log Odds			Mean SE of Log Odds			Coverage of 95% CI		
% Exposed	Drug	Separate[a]	Full[b]	Hierarchical	Separate[a]	Full[b]	Hierarchical	Separate[a]	Full[b]	Hierarchical
<5%	EFV	-0.019	0.013	-0.491	0.771	0.830	0.506	0.97	0.97	0.89
	ETR	-0.016	-0.002	0.053	0.761	0.800	0.481	0.97	0.97	0.98
	NVP	-0.075	-0.065	0.051	0.672	0.729	0.469	0.97	0.97	0.98
	FPV	0.038	0.053	0.001	0.797	0.855	0.449	0.96	0.96	0.99
	NFV	-0.103	-0.078	-0.006	0.493	0.594	0.408	0.97	0.97	0.99
5-15%	ABC	-0.021	-0.023	0.010	0.301	0.439	0.305	0.96	0.95	0.98
	RPV	-0.079	-0.050	0.031	0.437	0.520	0.417	0.96	0.95	0.96
	DRV	-0.034	-0.015	-0.004	0.286	0.389	0.339	0.95	0.95	0.96
>15%	3TC	-0.015	0.090	0.025	0.196	0.718	0.345	0.95	0.96	1.00
	FTC	0.012	-0.095	0.022	0.195	0.944	0.350	0.95	0.97	1.00
	TDF	0.014	0.178	0.023	0.195	0.939	0.346	0.95	0.96	0.99
	ZDV	-0.018	0.002	0.011	0.197	0.482	0.304	0.96	0.96	0.99
	ATV	-0.025	0.007	0.001	0.224	0.357	0.321	0.95	0.94	0.95
	LPV/r	-0.027	0.002	0.006	0.217	0.385	0.327	0.96	0.95	0.96

Abbreviations: ABC, Abacavir; ATV, Atazanavir; CI, confidence interval; DRV, Darunavir; EFV, Efavirenz; ETR, Etravirine; FPV, Fosamprenavir; FTC, Emtricitabine; LPV/r, Ritonavir-boosted Lopinavir; NFV, Nelfinavir; NVP, Nevirapine; RPV, Rilpivirine; SE, standard error; TDF, Tenofovir; ZDV, Zidovudine; 3TC, Lamivudine.

[a] 7.3% of the EFV models, 10.6% of the ETR models, 13.0% of the FPV models, 0.4% of the NFV models, 4.8% of the NVP models, and 0.1% of the RPV models did not converge or yielded unreasonable SEs. Results presented exclude these models.

[b] 21.1% of the full models did not converge. Results presented exclude these models.

Table A1.6 Scenario (i): bias, efficiency, and coverage of 95% confidence intervals across statistical approaches for the continuous outcome under scenario (i) (no true effects). Results are from 3,000 simulations each with sample size 1,000.

		Mean Bias in Difference			Mean SE of Difference			Coverage of 95% CI		
% Exposed	Drug	Separate	Full	Hierarchical	Separate	Full	Hierarchical	Separate	Full	Hierarchical
<5%	EFV	-0.006	-0.006	-0.005	0.298	0.313	0.280	0.95	0.95	0.97
	ETR	-0.001	-0.002	-0.002	0.238	0.248	0.231	0.95	0.95	0.96
	NVP	-0.000	-0.000	-0.001	0.204	0.221	0.210	0.95	0.95	0.96
	FPV	-0.004	-0.004	-0.004	0.256	0.273	0.249	0.95	0.95	0.97
	NFV	0.003	0.003	0.002	0.150	0.182	0.176	0.95	0.95	0.96
5-15%	ABC	0.002	0.003	0.003	0.096	0.139	0.132	0.95	0.95	0.96
	RPV	-0.000	-0.001	-0.001	0.135	0.161	0.157	0.95	0.96	0.96
	DRV	-0.001	-0.001	-0.001	0.091	0.123	0.121	0.95	0.95	0.96
>15%	3TC	-0.000	-0.007	-0.006	0.063	0.224	0.201	0.95	0.95	0.97
	FTC	-0.000	-0.006	-0.004	0.063	0.309	0.243	0.95	0.96	0.99
	TDF	-0.000	0.002	0.000	0.063	0.310	0.243	0.95	0.96	0.99
	ZDV	0.000	0.003	0.002	0.064	0.154	0.143	0.95	0.95	0.96
	ATV	0.000	-0.000	-0.000	0.072	0.114	0.112	0.95	0.95	0.95
	LPV/r	-0.001	-0.000	-0.001	0.070	0.123	0.120	0.95	0.95	0.96

Abbreviations: ABC, Abacavir; ATV, Atazanavir; CI, confidence interval; DRV, Darunavir; EFV, Efavirenz; ETR, Etravirine; FPV, Fosamprenavir; FTC, Emtricitabine; LPV/r, Ritonavir-boosted Lopinavir; NFV, Nelfinavir; NVP, Nevirapine; RPV, Rilpivirine; SE, standard error; TDF, Tenofovir; ZDV, Zidovudine; 3TC, Lamivudine.

Table A1.7 Scenario (ii): bias, efficiency, and coverage of 95% confidence intervals across statistical approaches for the continuous outcome under scenario (ii) (all protease inhibitors have a subtle effect on Bayley-III score). Results are from 3,000 simulations each with sample size 1,000.

		Mean Bias in Difference			Mean SE of Difference			Coverage of 95% CI		
% Exposed	Drug	Separate	Full	Hierarchical	Separate	Full	Hierarchical	Separate	Full	Hierarchical
<5%	EFV	0.228	0.007	0.006	0.301	0.313	0.280	0.88	0.95	0.97
	ETR	0.085	0.000	0.001	0.240	0.248	0.231	0.93	0.94	0.95
	NVP	0.213	0.008	0.007	0.206	0.221	0.210	0.82	0.96	0.97
	FPV	0.213	0.009	0.009	0.258	0.272	0.249	0.86	0.95	0.96
	NFV	0.219	0.009	0.009	0.151	0.182	0.176	0.70	0.95	0.95
5-15%	ABC	0.108	0.001	-0.000	0.097	0.139	0.132	0.80	0.95	0.96
	RPV	0.228	0.008	0.007	0.136	0.161	0.157	0.60	0.95	0.95
	DRV	0.214	0.009	0.009	0.092	0.123	0.121	0.35	0.95	0.95
>15%	3TC	-0.025	-0.020	-0.017	0.064	0.224	0.201	0.93	0.95	0.97
	FTC	0.017	-0.005	-0.008	0.064	0.309	0.243	0.94	0.96	0.99
	TDF	0.018	-0.016	-0.012	0.064	0.310	0.243	0.94	0.96	0.99
	ZDV	-0.011	-0.002	-0.003	0.064	0.154	0.143	0.95	0.95	0.97
	ATV	0.197	0.007	0.006	0.073	0.114	0.112	0.22	0.95	0.96
	LPV/r	0.196	0.009	0.008	0.071	0.123	0.120	0.20	0.95	0.96

Abbreviations: ABC, Abacavir; ATV, Atazanavir; CI, confidence interval; DRV, Darunavir; EFV, Efavirenz; ETR, Etravirine; FPV, Fosamprenavir; FTC, Emtricitabine; LPV/r, Ritonavir-boosted Lopinavir; NFV, Nelfinavir; NVP, Nevirapine; RPV, Rilpivirine; SE, standard error; TDF, Tenofovir; ZDV, Zidovudine; 3TC, Lamivudine.

Table A1.8 Scenario (iii.a): bias, efficiency, and coverage of 95% confidence intervals across statistical approaches for the continuous outcome under scenario (iii.a) (LPV/r has a moderate effect on Bayley-III score). Results are from 3,000 simulations each with sample size 1,000.

		Mean Bias in Difference			Mean SE of Difference			Coverage of 95% CI		
% Exposed	Drug	Separate	Full	Hierarchical	Separate	Full	Hierarchical	Separate	Full	Hierarchical
<5%	EFV	0.146	0.005	0.005	0.306	0.313	0.280	0.93	0.95	0.97
	ETR	0.132	0.012	0.011	0.244	0.248	0.231	0.92	0.95	0.97
	NVP	0.129	0.002	0.003	0.209	0.221	0.210	0.91	0.95	0.96
	FPV	0.119	-0.001	-0.022	0.262	0.272	0.249	0.93	0.95	0.96
	NFV	0.146	-0.002	-0.005	0.153	0.182	0.176	0.85	0.95	0.96
5-15%	ABC	0.110	0.003	0.007	0.099	0.139	0.132	0.79	0.94	0.95
	RPV	0.153	0.005	0.004	0.138	0.161	0.157	0.80	0.95	0.95
	DRV	0.169	0.004	0.001	0.093	0.123	0.121	0.56	0.95	0.95
>15%	3TC	-0.221	0.001	-0.003	0.065	0.224	0.201	0.07	0.95	0.97
	FTC	0.215	-0.004	-0.001	0.064	0.309	0.243	0.08	0.95	0.99
	TDF	0.216	0.000	-0.000	0.064	0.310	0.243	0.08	0.95	0.99
	ZDV	-0.226	-0.003	-0.004	0.065	0.154	0.143	0.07	0.95	0.96
	ATV	0.193	0.003	0.000	0.074	0.114	0.112	0.25	0.95	0.95
	LPV/r	0.000	0.002	0.012	0.070	0.123	0.120	0.95	0.95	0.95

Abbreviations: ABC, Abacavir; ATV, Atazanavir; CI, confidence interval; DRV, Darunavir; EFV, Efavirenz; ETR, Etravirine; FPV, Fosamprenavir; FTC, Emtricitabine; LPV/r, Ritonavir-boosted Lopinavir; NFV, Nelfinavir; NVP, Nevirapine; RPV, Rilpivirine; SE, standard error; TDF, Tenofovir; ZDV, Zidovudine; 3TC, Lamivudine.

Table A1.9 Scenario (iii.b): bias, efficiency, and coverage of 95% confidence intervals across statistical approaches for the continuous outcome under scenario (iii.b) (ABC has a moderate effect on Bayley-III score). Results are from 3,000 simulations each with sample size 1,000.

		Mean Bias in Difference			Mean SE of Difference			Coverage of 95% CI		
% Exposed	Drug	Separate	Full	Hierarchical	Separate	Full	Hierarchical	Separate	Full	Hierarchical
<5%	EFV	0.065	0.005	0.012	0.302	0.314	0.280	0.95	0.95	0.97
	ETR	0.057	-0.003	-0.001	0.241	0.248	0.231	0.95	0.96	0.97
	NVP	-0.035	0.003	0.010	0.207	0.221	0.210	0.95	0.95	0.96
	FPV	-0.063	0.004	0.008	0.259	0.273	0.249	0.94	0.95	0.97
	NFV	0.047	0.004	0.016	0.152	0.182	0.176	0.94	0.95	0.96
5-15%	ABC	-0.000	0.002	0.027	0.096	0.140	0.132	0.94	0.95	0.95
	RPV	0.052	0.003	0.009	0.137	0.161	0.157	0.93	0.94	0.95
	DRV	0.007	0.002	0.009	0.092	0.123	0.121	0.95	0.95	0.95
>15%	3TC	-0.117	0.004	-0.029	0.064	0.224	0.201	0.56	0.95	0.97
	FTC	0.114	0.009	0.002	0.064	0.309	0.244	0.57	0.94	0.99
	TDF	0.106	-0.005	-0.015	0.064	0.310	0.243	0.62	0.94	0.98
	ZDV	-0.036	0.000	0.008	0.065	0.154	0.143	0.91	0.95	0.96
	ATV	0.034	0.003	0.010	0.073	0.114	0.112	0.92	0.95	0.95
	LPV/r	0.059	0.005	0.017	0.071	0.123	0.120	0.87	0.95	0.95

Abbreviations: ABC, Abacavir; ATV, Atazanavir; CI, confidence interval; DRV, Darunavir; EFV, Efavirenz; ETR, Etravirine; FPV, Fosamprenavir; FTC, Emtricitabine; LPV/r, Ritonavir-boosted Lopinavir; NFV, Nelfinavir; NVP, Nevirapine; RPV, Rilpivirine; SE, standard error; TDF, Tenofovir; ZDV, Zidovudine; 3TC, Lamivudine.

Table A1.10 Scenario (iii.c): bias, efficiency, and coverage of 95% confidence intervals across statistical approaches for the continuous outcome under scenario (iii.c) (EFV has a moderate effect on Bayley-III score). Results are from 3,000 simulations each with sample size 1,000.

		Mean Bias in Difference			Mean SE of Difference			Coverage of 95% CI		
% Exposed	Drug	Separate	Full	Hierarchical	Separate	Full	Hierarchical	Separate	Full	Hierarchical
<5%	EFV	0.000	0.000	0.101	0.298	0.314	0.280	0.95	0.95	0.96
	ETR	0.005	-0.001	-0.021	0.238	0.248	0.231	0.95	0.95	0.96
	NVP	0.006	0.000	-0.013	0.205	0.221	0.210	0.95	0.95	0.96
	FPV	-0.003	-0.009	-0.005	0.256	0.273	0.249	0.95	0.95	0.97
	NFV	0.003	-0.002	-0.001	0.150	0.182	0.176	0.96	0.95	0.96
5-15%	ABC	0.003	-0.005	-0.004	0.096	0.140	0.132	0.96	0.95	0.96
	RPV	0.004	-0.002	-0.006	0.135	0.161	0.157	0.95	0.96	0.96
	DRV	0.008	0.001	0.004	0.091	0.123	0.121	0.96	0.96	0.96
>15%	3TC	0.009	-0.002	-0.005	0.064	0.224	0.201	0.95	0.95	0.97
	FTC	-0.010	-0.005	-0.005	0.063	0.309	0.244	0.95	0.96	0.99
	TDF	-0.010	0.000	-0.003	0.063	0.310	0.243	0.95	0.95	0.98
	ZDV	0.008	-0.003	-0.001	0.064	0.154	0.143	0.95	0.95	0.97
	ATV	0.009	0.001	0.004	0.072	0.114	0.112	0.95	0.95	0.96
	LPV/r	0.011	0.002	0.004	0.070	0.123	0.120	0.94	0.95	0.95

Abbreviations: ABC, Abacavir; ATV, Atazanavir; CI, confidence interval; DRV, Darunavir; EFV, Efavirenz; ETR, Etravirine; FPV, Fosamprenavir; FTC, Emtricitabine; LPV/r, Ritonavir-boosted Lopinavir; NFV, Nelfinavir; NVP, Nevirapine; RPV, Rilpivirine; SE, standard error; TDF, Tenofovir; ZDV, Zidovudine; 3TC, Lamivudine.

Table A2.1 The Mean Estimated Relative Excess Risk due to Interaction (RERI) and Empirical Coverage Rates of 95% Confidence Intervals for the RERI as Estimated From Generalized Estimating Equations (20 Clusters).[a]

			Log Binomial GEE			Poisson GEE		
				95% CI Coverage			95% CI Coverage	
Number of Clusters	Baseline Outcome Prevalence	RERI (RR_{01}, RR_{10}, RR_{11})	Mean $\widehat{RERI}$	Delta	MOVER	Mean $\widehat{RERI}$	Delta	MOVER
20	0.01	5 (2/3/9)	5.29	92.9	92.7	5.29	92.9	92.8
		10 (2/3/14)	10.60	91.9	90.5	10.60	91.8	90.5
	0.1	1 (1/1/2)	0.99	92.9	93.4	0.99	93.0	93.4
		1 (2/3/5)	1.00	92.4	92.7	1.00	92.1	92.6
		3 (1/2/5)	3.01	94.0	94.2	3.01	94.1	94.3
	0.2	1 (1/1/2)	1.01	92.8	93.0	1.01	92.9	92.9
		1 (1/2/3)	1.00	92.4	92.4	1.00	92.4	92.4
		1 (2/2/4)	1.00	91.6	91.7	1.00	91.6	91.7
	0.4	1 (1/1/2)	1.00	93.3	93.0	1.00	93.3	92.9
		0.5 (1/1/1.5)	0.50	92.5	92.9	0.50	92.6	92.9
	0.6	0.5 (1/1/1.5)	0.50	93.2	93.5	0.50	93.3	93.4

Abbreviations: CI, confidence interval; GEE, generalized estimating equations; MOVER, method of variance estimates recovery; RERI, relative excess risk due to interaction; RR_{01}, relative risk of outcome in the group unexposed to X_1 and exposed to X_2 as compared to the doubly unexposed group; RR_{10}, relative risk of outcome in the group exposed to X_1 and unexposed to X_2 as compared to the doubly unexposed group; RR_{11}, relative risk of outcome in the doubly exposed group as compared to the doubly unexposed group.

[a] Generalized estimating equations estimated specifying an exchangeable working correlation, log link, and either binomial or Poisson distribution, with robust variance estimates. Results are based on 2,000 simulated datasets.

Table A2.2 The Mean Estimated Relative Excess Risk due to Interaction (RERI) and Empirical Coverage Rates of 95% Confidence Intervals for the RERI as Estimated From Generalized Estimating Equations (50 Clusters).[a]

			Log Binomial GEE			Poisson GEE		
				95% CI Coverage			95% CI Coverage	
Number of Clusters	Baseline Outcome Prevalence	RERI (RR_{01}, RR_{10}, RR_{11})	Mean $\widehat{RERI}$	Delta	MOVER	Mean $\widehat{RERI}$	Delta	MOVER
50	0.01	5 (2/3/9)	5.12	93.6	94.4	5.12	93.6	94.4
		10 (2/3/14)	10.36	93.2	93.6	10.36	93.2	93.6
	0.1	1 (1/1/2)	1.00	94.7	94.8	1.00	94.7	94.8
		1 (2/3/5)	0.99	94.6	94.4	0.99	94.7	94.4
		3 (1/2/5)	3.01	94.2	94.4	3.01	94.3	94.4
	0.2	1 (1/1/2)	1.00	94.4	94.5	1.00	94.3	94.6
		1 (1/2/3)	1.00	94.7	94.7	1.00	94.7	94.7
		1 (2/2/4)	0.99	93.6	94.0	0.99	93.7	94.0
	0.4	1 (1/1/2)	1.00	93.5	93.7	1.00	93.6	93.7
		0.5 (1/1/1.5)	0.50	94.0	94.5	0.50	94.1	94.3
	0.6	0.5 (1/1/1.5)	0.50	94.7	94.3	0.50	94.7	94.4

Abbreviations: CI, confidence interval; GEE, generalized estimating equations; MOVER, method of variance estimates recovery; RERI, relative excess risk due to interaction; RR_{01}, relative risk of outcome in the group unexposed to X_1 and exposed to X_2 as compared to the doubly unexposed group; RR_{10}, relative risk of outcome in the group exposed to X_1 and unexposed to X_2 as compared to the doubly unexposed group; RR_{11}, relative risk of outcome in the doubly exposed group as compared to the doubly unexposed group.

[a] Generalized estimating equations estimated specifying an exchangeable working correlation, log link, and either binomial or Poisson distribution, with robust variance estimates. Results are based on 2,000 simulated datasets.

Table A2.3 The Mean Estimated Relative Excess Risk due to Interaction (RERI) and Empirical Coverage Rates of 95% Confidence Intervals for the RERI as Estimated From Naïve Log Binomial and Naïve Logistic Regression Models (20 Clusters).[a]

			Naive Log Binomial			Naive Logistic		
				95% CI Coverage			95% CI Coverage	
Number of Clusters	Baseline Outcome Prevalence	RERI ($RR_{01}/RR_{10}/RR_{11}$)	Mean $\widehat{RERI}$	Delta	MOVER	Mean $\widehat{RERI}$	Delta	MOVER
20	0.01	5 (2/3/9)	5.29	95.5	94.9	6.24	97.4	92.6
		10 (2/3/14)	10.60	94.1	93.5	13.17	97.8	85.2
	0.1	1 (1/1/2)	1.00	95.3	95.2	1.28	91.6	90.4
		1 (2/3/5)	1.00	94.4	94.7	4.05	28.5	23.5
		3 (1/2/5)	3.01	95.9	96.4	6.99	4.1	2.3
	0.2	1 (1/1/2)	1.01	95.6	95.7	1.73	53.7	48.1
		1 (1/2/3)	1.00	94.8	94.9	3.45	9.1	6.4
		1 (2/2/4)	1.00	94.1	94.0	12.07	0.0	0.0
	0.4	1 (1/1/2)	1.00	95.2	95.5	5.12	0.0	0.0
		0.5 (1/1/1.5)	0.50	95.0	95.4	1.28	29.4	25.6
	0.6	0.5 (1/1/1.5)	0.50	94.8	95.1	5.26	0.0	0.0

Abbreviations: CI, confidence interval; MOVER, method of variance estimates recovery; RERI, relative excess risk due to interaction; RR_{01}, relative risk of outcome in the group unexposed to X_1 and exposed to X_2 as compared to the doubly unexposed group; RR_{10}, relative risk of outcome in the group exposed to X_1 and unexposed to X_2 as compared to the doubly unexposed group; RR_{11}, relative risk of outcome in the doubly exposed group as compared to the doubly unexposed group.

[a]Results are based on 2,000 simulated datasets.

Table A2.4 The Mean Estimated Relative Excess Risk due to Interaction (RERI) and Empirical Coverage Rates of 95% Confidence Intervals for the RERI as Estimated From Naïve Log Binomial and Naïve Logistic Regression Models (50 Clusters).[a]

			Naive Log Binomial			Naive Logistic		
				95% CI Coverage			95% CI Coverage	
Number of Clusters	Baseline Outcome Prevalence	RERI ($RR_{01}/RR_{10}/RR_{11}$)	Mean $\widehat{RERI}$	Delta	MOVER	Mean $\widehat{RERI}$	Delta	MOVER
50	0.01	5 (2/3/9)	5.13	94.6	95.5	6.05	97.1	93.4
		10 (2/3/14)	10.36	94.0	94.9	12.85	98.1	87.1
	0.1	1 (1/1/2)	0.99	95.6	95.6	1.27	91.5	90.1
		1 (2/3/5)	0.99	95.0	95.2	4.02	23.9	18.8
		3 (1/2/5)	3.01	95.3	95.2	6.96	2.8	1.6
	0.2	1 (1/1/2)	1.00	94.6	94.8	1.71	52.2	47.9
		1 (1/2/3)	1.00	95.0	95.5	3.44	7.4	5.2
		1 (2/2/4)	0.99	94.7	94.8	12.01	0.0	0.0
	0.4	1 (1/1/2)	1.00	94.7	94.7	5.13	0.0	0.0
		0.5 (1/1/1.5)	0.50	94.7	94.7	1.28	26.8	23.8
	0.6	0.5 (1/1/1.5)	0.50	94.8	95.1	5.21	0.0	0.0

Abbreviations: CI, confidence interval; MOVER, method of variance estimates recovery; RERI, relative excess risk due to interaction; RR_{01}, relative risk of outcome in the group unexposed to X_1 and exposed to X_2 as compared to the doubly unexposed group; RR_{10}, relative risk of outcome in the group exposed to X_1 and unexposed to X_2 as compared to the doubly unexposed group; RR_{11}, relative risk of outcome in the doubly exposed group as compared to the doubly unexposed group.

[a]Results are based on 2,000 simulated datasets.

Table A2.5 The Mean Estimated Relative Excess Risk due to Interaction (RERI) and Empirical Coverage Rates of 95% Confidence Intervals for the RERI as Estimated From Naïve Log Binomial and Naïve Logistic Regression Models (275 Clusters).[a]

			Naive Log Binomial			Naive Logistic		
				95% CI Coverage			95% CI Coverage	
Number of Clusters	Baseline Outcome Prevalence	RERI ($RR_{01}/RR_{10}/RR_{11}$)	Mean $\widehat{RERI}$	Delta	MOVER	Mean $\widehat{RERI}$	Delta	MOVER
275	0.01	5 (2/3/9)	5.25	95.0	94.9	6.21	97.2	92.4
		10 (2/3/14)	10.60	95.5	93.7	13.18	98.3	85.1
	0.1	1 (1/1/2)	1.00	95.0	95.1	1.28	92.1	90.5
		1 (2/3/5)	0.98	95.5	95.6	4.01	27.3	22.5
		3 (1/2/5)	3.00	95.5	95.3	6.91	3.4	2.2
	0.2	1 (1/1/2)	1.00	95.6	95.6	1.70	54.1	48.3
		1 (1/2/3)	0.99	95.1	95.3	3.43	8.3	5.5
		1 (2/2/4)	1.00	95.4	95.5	12.04	0.0	0.0
	0.4	1 (1/1/2)	1.00	95.4	95.7	5.12	0.0	0.0
		0.5 (1/1/1.5)	0.50	95.2	95.1	1.28	26.2	22.7
	0.6	0.5 (1/1/1.5)	0.50	95.2	95.4	5.22	0.0	0.0

Abbreviations: CI, confidence interval; MOVER, method of variance estimates recovery; RERI, relative excess risk due to interaction; RR_{01}, relative risk of outcome in the group unexposed to X_1 and exposed to X_2 as compared to the doubly unexposed group; RR_{10}, relative risk of outcome in the group exposed to X_1 and unexposed to X_2 as compared to the doubly unexposed group; RR_{11}, relative risk of outcome in the doubly exposed group as compared to the doubly unexposed group.

[a]Results are based on 2,000 simulated datasets.

Table A2.6 The Median Width of 95% Confidence Intervals for the Relative Excess Risk due to Interaction (RERI) as Estimated From Naïve Log Binomial and GEE Log Binomial Models (20 Clusters).[a]

Number of Clusters	Baseline Outcome Prevalence	RERI ($RR_{01}/RR_{10}/RR_{11}$)	Delta Method		MOVER Method	
			Marginal Log Binomial	Naive Log Binomial	Marginal Log Binomial	Naive Log Binomial
20	0.01	5 (2/3/9)	7.66	8.18	9.05	9.70
		10 (2/3/14)	11.37	11.86	12.57	13.17
	0.1	1 (1/1/2)	1.13	1.21	1.17	1.25
		1 (1/2/3)	1.34	1.41	1.38	1.45
		1 (2/3/5)	1.58	1.67	1.63	1.71
		3 (1/2/5)	1.52	1.61	1.56	1.65
	0.2	1 (1/1/2)	0.75	0.80	0.77	0.81
		1 (1/2/3)	0.80	0.84	0.81	0.85
		1 (2/2/4)	0.80	0.85	0.81	0.86
	0.4	1 (1/1/2)	0.40	0.42	0.41	0.42
		0.5 (1/1/1.5)	0.44	0.46	0.44	0.46
		0.5 (1/1/1.5)	0.43	0.45	0.43	0.46
	0.6	0.5 (1/1/1.5)	0.25	0.26	0.25	0.27

Abbreviations: CI, confidence interval; GEE, generalized estimating equations; MOVER, method of variance estimates recovery; RERI, relative excess risk due to interaction; RR_{01}, relative risk of outcome in the group unexposed to X_1 and exposed to X_2 as compared to the doubly unexposed group; RR_{10}, relative risk of outcome in the group exposed to X_1 and unexposed to X_2 as compared to the doubly unexposed group; RR_{11}, relative risk of outcome in the doubly exposed group as compared to the doubly unexposed group.

[a] Generalized estimating equations estimated specifying an exchangeable working correlation, log link, and binomial distribution, with robust variance estimates. Results are based on 2,000 simulated datasets.

Table A2.7 The Median Width of 95% Confidence Intervals for the Relative Excess Risk due to Interaction (RERI) as Estimated From Naïve Log Binomial and GEE Log Binomial Models (50 Clusters).[a]

Number of Clusters	Baseline Outcome Prevalence	RERI ($RR_{01}/RR_{10}/RR_{11}$)	Delta Method		MOVER Method	
			Marginal Log Binomial	Naive Log Binomial	Marginal Log Binomial	Naive Log Binomial
50	0.01	5 (2/3/9)	7.61	7.78	8.79	9.05
		10 (2/3/14)	11.07	11.46	12.24	12.57
	0.1	1 (1/1/2)	1.14	1.17	1.18	1.21
		1 (1/2/3)	1.32	1.36	1.36	1.40
		1 (2/3/5)	1.59	1.61	1.63	1.65
		3 (1/2/5)	1.52	1.55	1.55	1.59
	0.2	1 (1/1/2)	0.75	0.77	0.76	0.78
		1 (1/2/3)	0.80	0.81	0.81	0.82
		1 (2/2/4)	0.80	0.82	0.81	0.82
	0.4	1 (1/1/2)	0.40	0.41	0.40	0.41
		0.5 (1/1/1.5)	0.44	0.44	0.44	0.44
		0.5 (1/1/1.5)	0.43	0.44	0.43	0.44
	0.6	0.5 (1/1/1.5)	0.25	0.26	0.25	0.26

Abbreviations: CI, confidence interval; GEE, generalized estimating equations; MOVER, method of variance estimates recovery; RERI, relative excess risk due to interaction; RR_{01}, relative risk of outcome in the group unexposed to X_1 and exposed to X_2 as compared to the doubly unexposed group; RR_{10}, relative risk of outcome in the group exposed to X_1 and unexposed to X_2 as compared to the doubly unexposed group; RR_{11}, relative risk of outcome in the doubly exposed group as compared to the doubly unexposed group.

[a] Generalized estimating equations estimated specifying an exchangeable working correlation, log link, and binomial distribution, with robust variance estimates. Results are based on 2,000 simulated datasets.

Table A2.8 The Median Width of 95% Confidence Intervals for the Relative Excess Risk due to Interaction (RERI) as Estimated From Naïve Log Binomial and GEE Log Binomial Models (275 Clusters).[a]

Number of Clusters	Baseline Outcome Prevalence	RERI ($RR_{01}/RR_{10}/RR_{11}$)	Delta Method		MOVER Method	
			Marginal Log Binomial	Naive Log Binomial	Marginal Log Binomial	Naive Log Binomial
275	0.01	5 (2/3/9)	8.00	8.04	9.38	9.39
		10 (2/3/14)	11.59	11.66	12.80	12.81
	0.1	1 (1/1/2)	1.19	1.20	1.23	1.24
		1 (1/2/3)	1.37	1.39	1.41	1.43
		1 (2/3/5)	1.63	1.64	1.68	1.69
		3 (1/2/5)	1.57	1.58	1.61	1.62
	0.2	1 (1/1/2)	0.78	0.79	0.79	0.80
		1 (1/2/3)	0.82	0.83	0.83	0.84
		1 (2/2/4)	0.83	0.83	0.84	0.84
	0.4	1 (1/1/2)	0.41	0.41	0.42	0.42
		0.5 (1/1/1.5)	0.45	0.45	0.45	0.45
		0.5 (1/1/1.5)	0.44	0.45	0.45	0.45
	0.6	0.5 (1/1/1.5)	0.26	0.26	0.26	0.26

Abbreviations: CI, confidence interval; GEE, generalized estimating equations; MOVER, method of variance estimates recovery; RERI, relative excess risk due to interaction; RR_{01}, relative risk of outcome in the group unexposed to X_1 and exposed to X_2 as compared to the doubly unexposed group; RR_{10}, relative risk of outcome in the group exposed to X_1 and unexposed to X_2 as compared to the doubly unexposed group; RR_{11}, relative risk of outcome in the doubly exposed group as compared to the doubly unexposed group.

[a] Generalized estimating equations estimated specifying an exchangeable working correlation, log link, and binomial distribution, with robust variance estimates. Results are based on 2,000 simulated datasets.

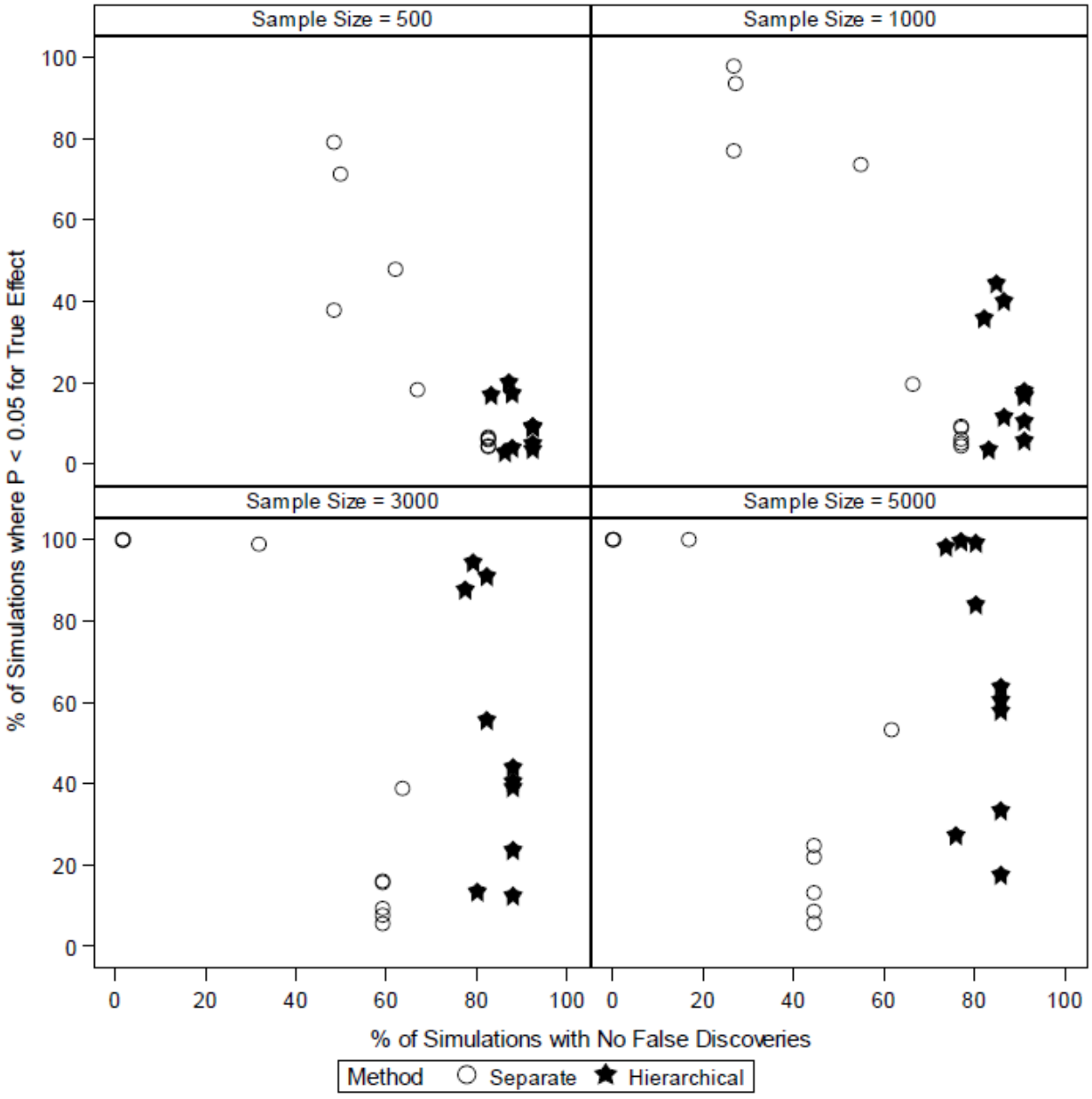

Figure A1.1 The percent of simulations with no false discoveries versus the power to detect the true effects of antiretroviral exposures on preterm birth. Markers represent the drugs with true effects under different exposure-outcome scenarios. Results are based on 3,000 simulations.

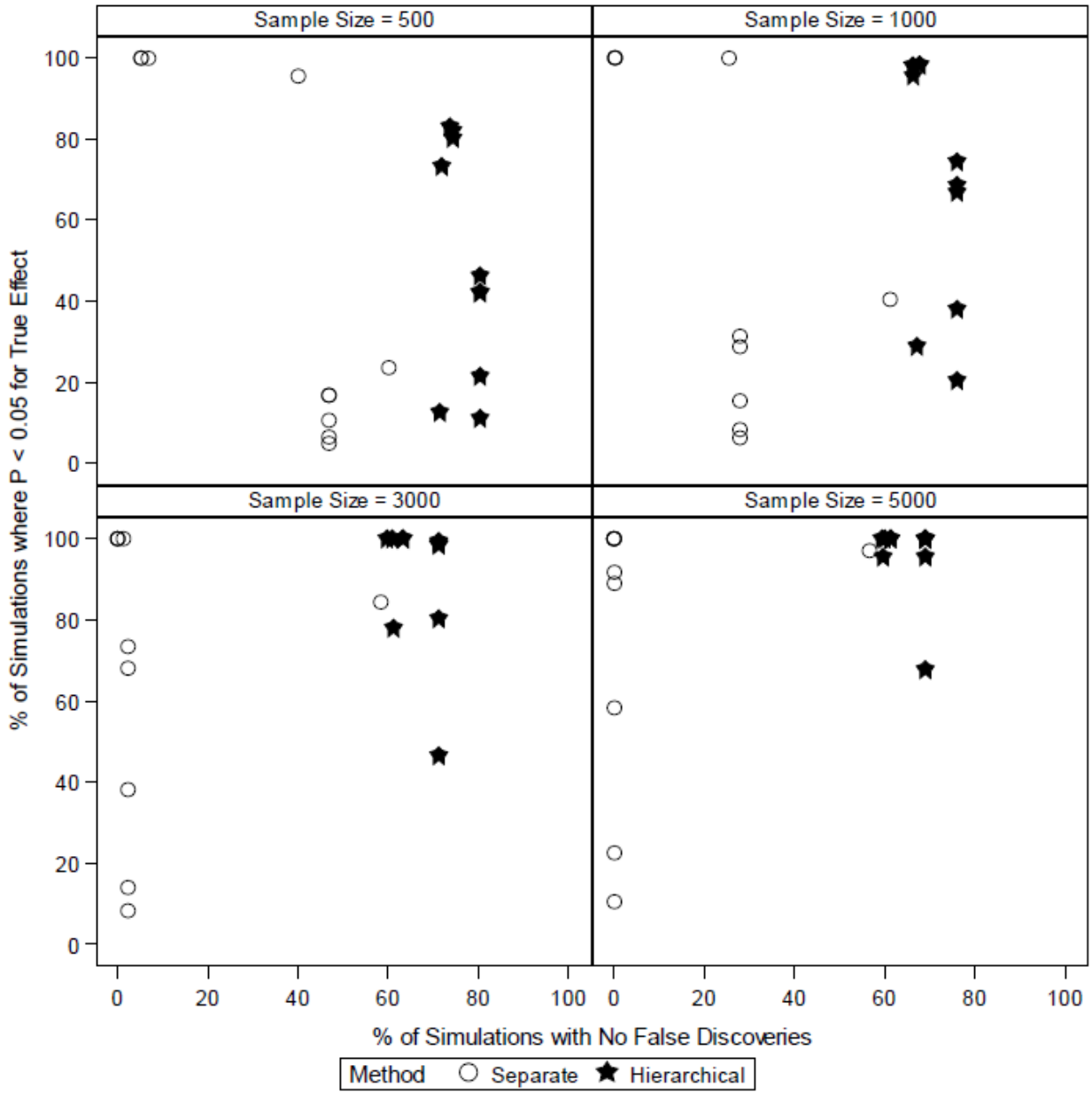

Figure A1.2 The percent of simulations with no false discoveries versus the power to detect the true effects of antiretroviral exposures on standardized Bayley-III score. Markers represent the drugs with true effects under different exposure-outcome scenarios. Results are based on 3,000 simulations.

References

Aanerud M, Carsin A, Sunyer J, et al. Interaction between asthma and smoking increases the risk of adult airway obstruction. *Eur Respir J*. 2015;45:586-588.

Abers MS, Shandera WX, Kass JS. Neurological and psychiatric adverse effects of antiretroviral drugs. *CNS Drugs* 2014; **28**:131-45.

Aragaki CC, Greenland S, Probst-Hensch N, et al. Hierarchical modeling of gene-environment interaction: Estimating NAT2* genotype specific dietary effects on adenomatous polyps. *Cancer Epidemiology, Biomarkers & Prevention* 1997;**6**:307-314.

Aschengrau A, Seage III GR. *Essentials of Epidemiology in Public Health.* 3rd ed. Sudbury, MA: Jones and Bartlett Publishers; 2014.

Assman SF, Hosmer DW, Lemeshow S, et al. Confidence intervals for measures of interaction. *Epidemiology.* 1996;7(3):286-290.

Auer PL, Lettre G. Rare variant association studies: challenges and opportunities. *Genome Medicine.* 2015; 7-16.

Barros AJ, Hirakata VN. Alternative for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Med Res Methodol.* 2003;3:21.

Bermedo-Carrasco S, Pena-Sanchez JN, Lepnurm R, et al. Inequities in cervical cancer screening among Colombian women: a multilevel analysis of a nationwide survey. *Cancer Epidemiology*. 2015;39:229-236.

Bisio F, Nicco E, Calzi A, et al. Pregnancy outcomes following exposure to efavirenz-based antiretroviral therapy in the Republic of Congo. *New Microbiologica* 2015;**38**:185-192.

Brenner D, Brennan P, Boffetta P, et al. Hierarchical modeling identifies novel lung cancer susceptibility variants in inflammation pathways among 10,140 cases and 11,012 controls. *Hum Genet* 2013;**132**:579-589.

Bürkner, PC (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1-28.<doi:10.18637/jss.v080.i01>

Caniglia EC, Patel K, Huo Y, et al. Atazanavir exposure in utero and neurodevelopment in infants: a comparative safety study. *AIDS*. 2016;**30**:1267-1278.

Capanu M, Orlow I, Berwick M, Hummer AJ, Thomas DC, Begg CB. The use of hierarchical models for estimating relative risks of individual genetic variants: An application to a study of melanoma. *Stat Med* 2008;**27**:1973-1992.

Capanu M, Begg C. Hierarchical modeling for estimating relative risks of rare genetic variants: properties of the pseudo-likelihood method. *Biometrics* 2011;**67**:371-380.

Carter RE, Lipsitz SR, Tilley BC. Quasi-likelihood estimation for relative risk regression models. *Biostatistics*. 2005;6(1):39-44.

CDC. Achievements in public health: reduction in perinatal transmission of HIV infection – United States, 1985-2005. *MMWR Morb Mortal Wkly Rep.* 2006; **55**:592-597.

Chen Y, Graziano JH, Parvez F, et al. Modification of risk of arsenic-induced skin lesions by sunlight exposure, smoking, and occupational exposures in Bangledesh. *Epidemiology*. 2006;17(4):459-467.

Chiu YH, Karmon AE, Gaskins AJ, et al. Serum omega-3 fatty acids and treatment outcomes among women undergoing assisted reproduction. *Hum Reprod.* 2018;33(1):156-165.

Chu H, Cole SR. Estimation of risk ratios in cohort studies with common outcomes: A Bayesian approach. *Epidemiology*. 2010;21(6):855-862.

Chung Y, Rabe-Hesketh S, Dorie V, et al. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*. 2013;78(4):685-709.

Cihlar T, Ray A. Nucleoside and nucleotide HIV reverse transcriptase inhibitors: 25 years after Zidovudine. *Antiviral Res* 2010; **85**: 39-58.

Conti DV, Witte JS. Hierarchical modeling of linkage disequilibrium: Genetic structure and spatial relations. *Am J Hum Genet* 2003;**72**:351-363.

Cook MB, Wild CP, Forman D. A systematic review and meta-analysis of the sex ratio for Barrett's esophagus, erosive reflux disease, and nonerosive reflux disease. *Am J Epidemiol.* 2005;162(11):1050-1061.

Cote HCF, Brumme ZL, Craib KJP et al. Changes in mitochondrial DNA as a marker of nucleoside toxicity in HIV-infected patients. *N Engl J Med* 2002; **346**:811-820.

Cotter AM, Garcia AG, Duthely ML, et al. Is antiretroviral therapy during pregnancy associated with an increased risk of preterm delivery, low birth weight, or stillbirth? *J Infect Dis.* 2006;**193**:1195-1201.

Crump C, Sundquist J, Winkleby MA, et al. Height, weight, and aerobic fitness in relation to risk of atrial fibrillation. [published online ahead of print June 21, 2017] *Am J Epidemiol*. (doi: 10.1093/aje/kwx255).

De Bethune MP. Non-nucleoside reverse transcriptase inhibitors (NNRTIs), their discovery, development, and use in the treatment of HIV-1 infection: a review of the last 20 years (1989-2009). *Antiviral Res* 2010; **85**: 75-90.

Deddens JA, Peterson MR, Lei X. Estimation of prevalence ratios when PROC GENMOD does not converge. (Paper 270-28). In: Proceedings of the 28th Annual SAS Users Group International Conference. Cary, NC: SAS Institute, Inc., 2003.

Dey R, Lee S (2017). SPAtest: Score Test Based on Saddlepoint Approximation. R package version 2.0.2. URL: https://CRAN.R-project.org/package=SPAtest

Dey R, Schmidt EM, Abecasis GR, Lee S. A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *Am J Human Genet*. 2017(101): 1-13.

Dupont C, Winer N, Rabilloud M, et al. Multifaceted intervention to improve obstetric practices: the OPERA cluster-randomized controlled trial. *Eur J Obstet Gynecol Reprod Biol*. 2017;215:206-212.

Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993(80):27-38.

Fitzmaurice GM, Lipsitz SR, Arriage A, et al. Almost efficient estimation of relative risk regression. *Biostatistics*. 2014;15(4):745-756.

Fowler MG, Qin M, Fiscus SA, et al. Benefits and risk of antiretroviral therapy for perinatal HIV prevention. *N Engl J Med*. 2016;375(18):1726-1737.

Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*. 2006;1(3):515-533.

Gemperli A, Vounatsou P, Kleinschmidt I, et al. Spatial patterns of infant mortality in Mali: the effect of malaria endemicity. *Am J Epidemiol.* 2004;159(1):64-72.

Goyette MS, Mutiti PM, Bukusi D, et al. HIV assisted partner services among those with and without a history of intimate partner violence in Kenya. *J Acquir Immune Defic Syndr.* 2018 Feb 5. [Epub ahead of print]

Greenland S. A semi-bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. *Stat Med.* 1992;**11**:219-230.

Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-bayes regression. *Stat Med.* 1993;**12**:717-736.

Greenland S. Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models in epidemiologic analyses. *Stat Med.* 1997; **16**:515-526.

Greenland S. When should epidemiologic regressions use random coefficients? *Biometrics.* 2000;**56**:915-921.

Griner R, Williams PL, Read JS, et al. In utero and postnatal exposure to antiretrovirals among HIV-exposed but uninfected children in the United States. *AIDS Patient Care STDS* 2011;**25**:385-394.

Grosch-Woerner I, Puch K, Maier RF, et al. Increased rate of prematurity associated with antenatal antiretroviral therapy in a German/Austrian cohort of HIV-1 infected women. *HIV Med* 2008;**9**:6-13.

Gustavsson J, Mehlig K, Leander K, et al. FTO gene variation, macronutrient intake and coronary heart disease risk: a gene-diet interaction analysis. *Eur J Nutr.* 2016;55(1):247-255.

Hecker J, Ruczinski I, Cho M, et al. A universal and nearly optimal permutation testing approach for association analysis in whole-genome sequencing studies. 2018. [Manuscript in preparation.]

Hagihara A, Onozuka D, Ono J, et al. Age x gender interaction effect on resuscitation outcomes in patients with out-of-hospital cardiac arrest. *Am J Cardiol.* 2017;120(3):387-392.

Hajat A, Allison M, Diez-Roux AV, et al. Long-term exposure to air pollution and markers of inflammation, coagulation, and endothelial activation: a repeat measures analysis in the multi-ethnic study of atherosclerosis (MESA). *Epidemiology*. 2015;26(3):310-320.

Hosmer DW, Lemeshow S. Confidence interval estimation of interaction. *Epidemiology*. 1992;3(5):452-456.

Hung RJ, Brennan P, Malaveille C, et al. Using hierarchical modeling in genetic association studies with multiple markers: Application to a case-control study of bladder cancer. *Cancer Epidemiology, Biomarkers & Prevention* 2004;**13**:1013-1021.

Jabbarpoor Bonyadi MH, Yaseri M, Bonyadi M, et al. Association of combined complement factor H Y402H and ARMS/LOC387715 A69S polymorphisms with age-related macular degeneration: a meta-analysis. *Curr Eye Res.* 2016;41(12):1519-1525.

Jabbarpoor Bonyadi MH, Yaseri M, Bonyadi M, et al. Association of combined cigarette smoking and ARMS2/LOC387715 A69S polymorphisms with age-related macular degeneration: A meta-analysis. *Ophthalmic Genet*. 2017;38(4):308-313.

Jian S, Su-Mei N, Xue C, et al. Association and interaction between triglyceride-glucose index and obesity on risk of hypertension in middle-aged and elderly adults. *Clin Exp Hypertens.* 2017;39(8):732-739.

Kalkut G. Antiretroviral therapy: an update for the non-AIDS specialist. *Curr Opin Oncol* 2005; **17**:479-84.

Kloog I, Melley SJ, Coull BA, et al. Using satellite-based spatiotemporal resolved air temperature exposure to study the association between ambient air temperature and birth outcomes in Massachusetts. *Environ Health Perspect.* 2015;123(10):1053-1058.

Knol MJ, Vandenbroucke JP, Scott P, et al. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. *Am J Epidemiol*. 2008;168(9):1073-1081.

Knol MJ, Egger M, Scott P, Geerlings MI, Vandenbroucke JP. When one depends on the other: reporting of interaction in case-control and cohort studies. *Epidemiology.* 2009;20(2):161-166.

Knol MJ, VanderWeele TJ, Groenwold RHH, et al. Estimating measures of interaction on an additive scale for preventive exposures. *Eur J Epidemiol.* 2011;26:433-438.

Knol M, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. *Int J Epidemiol*. 2012;41:514-520.

Koss CA, Natureeba P, Plenty A, et al. Risk factors for preterm birth among HIV-infected pregnant Ugandan women randomized to lopinavir/ritonavir- or efavirenz-based antiretroviral therapy. . *J Acquir Immune Defic Syndr*. 2014;**67**:128-135

Lee S, Abecasis GR, Boehnke M, et al. Rare-variant association analysis: study designs and statistical tests. *Am J Human Genet.* 2014(95): 5-23.

Li B, Leal SM .Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Human Genet.* 2008(83): 311-321.

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13-22.

Lin Y, MacLain AC, Probst JC, et al. Health-related quality of life among adults 65 years and older in the United States, 2011-2012: a multilevel small area estimation approach. *Ann Epidemiol.* 2017;27:52-58.

Lipsitz SR, Fitzmaurice GM, Arriaga A, et al. Using the jackknife for estimation in log link Bernoulli regression models. *Stat Med*. 2015;34(3):444-453.

Ma C, Blackwell T, Boehnke M, Scott LJ. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic Epidemiology.* 2013(37):539-550.

Madden JV, Flatley CJ, Kumar S. Term small-for-gestational age infants from low risk women are at significantly greater risk of adverse neonatal outcomes. *Am J Obstet Gynecol*. 2018; doi:10.1016/j.ajog.2018.02.008. [Epub ahead of print.]

Mao G, Massa Nachman R, Sun Q, et al. Individual and joint effects of early-life ambient $PM_{2.5}$ exposure and maternal pre-pregnancy obesity on childhood overweight or obesity. *Environ Health Perspect.* 2017;125(6):067005.

Marschner IC, Gillett A. Relative risk regression: reliable and flexible methods for log-binomial models. *Biostatistics.* 2012;13(1):179-192.

McNutt L, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol*. 2003;157(10):940-943.

Meng XY, Zhou Y, Zhang J, et al. The association and interaction analysis of hypertension and diabetes mellitus on diastolic heart failure in a high-risk population. *Int J Clin Exp Med.* 2015;8(11):21311-21318.

Menvielle G, Fayosse A, Radoi L, et al. The joint effect of asbestos exposure, tobacco smoking and alcohol drinking on laryngeal cancer risk: evidence from the Fresh population-based case-control study, ICARE. *J Occup Environ Med.* 2016; 73(1):28-33.

Mesfin YM, Kibret KT, Taye A. Is protease inhibitors based antiretroviral therapy during pregnancy associated with an increased risk of preterm birth? Systematic review and a meta-analysis. *Reprod Health* 2016 Apr 5;13:30.

Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res.* 2007(615):28-56.

Neuhaus JM. Statistical methods for longitudinal and clustered designs with binary responses. *Stat Methods Med Res*. 1992;1(3):249-273.

Oh HY, Kim MK, Seo SS, et al. Association of combined tobacco smoking and oral contraceptive use with cervical intraepithelial neoplasia 2 or 3 in Korean women. *J Epidemiol.* 2016;26(1):22-29.

Pedroza C, Thanh Truong VT. Performance of models for estimating the absolute risk difference in multicenter trials with binary outcome. *BMC Med Res Methodol*. 2016;16:113.

Perry M, Taylor GP, Sabin CA, et al. Lopinavir and atazanavir in pregnancy: comparable infant outcomes, virological efficacies and preterm delivery rates. *HIV Med* 2016;**17**:28-35.

Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*.197 9(66):403-411.

Prokopenko et al., 2018 [manuscript in preparation]

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Raifman J, Moscoe E, Bryn AS, et al. Difference-in-differences analysis of the association between state same-sex marriage policies and adolescent suicide attempts. *JAMA Pediatr.* 2017;171(4):350-356.

Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDGene) study design. *COPD.* 2010;7(1):32-43.

Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *Am J Epidemiol.* 1980;112(4):467-470

Rothman KJ. Writing for Epidemiology. *Epidemiology*. 1998;9(3):333-337.

Rothman KJ, Greenland S, and Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins, Publishers; 2008.

SAS/STAT(R) 9.22 User's Guide. The GLIMMIX Procedure, Model or Integral Approximation.https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_glimmix_a0000001430.htm. Accessed March 5, 2018.

Sibiude J, Warszawski J, Tubiana R, et al. Premature delivery in HIV-infected women starting protease inhibitor therapy during pregnancy: role of the ritonavir boost? *Clin Infect Dis.* 2012; **54**: 1348-1360.

Silverman EK, Chapman HA, Drazen JM, Weiss ST, Rosner B, Campbell EJ, et al. Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease. Risk to relatives for airflow obstruction and chronic bronchitis. *Am J Respir Crit Care Med*. 1998;157: 1770–1778. doi:10.1164/ajrccm.157.6.9706014

Simons CC, Schouten LJ, Godschalk RW, et al. Energy restriction at young age, genetic variants in the insulin-like growth factor pathway and colorectal cancer risk in the Netherlands Cohort Study. *Int J Cancer*. 2017;140(2):272-284.

Smith C, Weinberg A, Forster JE et al. Maternal lopinavir/ritonavir is associated with fewer adverse events in infants than nelfinavir or atazanavir. *Infect Dis Obstet Gynecol* 2016; 9848041.

Spiegelman D, Hertzmark E. Easy SAS Calculations for Risk or Prevalence Ratios and Differences. *Am J Epidemiol.* 2005;162(3):199-200.

Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.17.3. http://mc-stan.org/.

Stein JH. Dyslipidemia in the era of HIV protease inhibitors. *Prog Cardiovasc Dis.* 2003; **45**(4):293-304.

Suksomboon N, Poolsup N, Ket-aim S. Systematic review of the efficacy of antiretroviral therapies for reducing the risk of mother-to-child transmission of HIV infection. *J Clin Pharm Ther* 2007; **32**:293-311.

Tassiopoulos K, Williams PL, Seage GR 3rd, Crain M, Oleske J, Farley J, International Maternal Pediatric Adolescent AIDS Clinical Trials 219C Team. Association of hypercholesterolemia incidence with antiretroviral treatment, including protease inhibitors, among perinatally HIV-infected children. *J Acquir Immune Defic Syndr.* 2008;**47**(5):607-14.

Timpka S, Stuart JJ, Tanz LJ, et al. Lifestyle in progression from hypertensive disorders of pregnancy to chronic hypertension in Nurses' Health Study II: observational cohort study. *BMJ*. 2017; 358:j3024.

Torman VBL, Camey SA. Bayesian models as a unified approach to estimate relative risk (or prevalence ratio) in binary and polytomous outcomes. *Emerging Themes in Epidemiology.* 2015;12:8.

Tsai A, Weiser SD, Dilworth SE, et al. Violent victimization, mental health and service utilization outcomes in a cohort of homeless and unstably housed women living with or at risk of becoming infected with HIV. *Am J Epidemiol.* 2015;181(10):817-826.

Tuomala RE, Shapiro DE, Mofenson LM, et al. Antiretroviral therapy during pregnancy and the risk of adverse outcome. *N Engl J Med* 2002;**346**:1863-1870.

U.S. Department of Health and Human Services. AIDSInfo -- Recommendations for Use of Antiretroviral Drugs in Pregnant HIV-1-Infected Women for Maternal Health and Interventions to Reduce Perinatal HIV Transmission in the United States. https://aidsinfo.nih.gov/guidelines/html/3/perinatal-guidelines/204/nevirapine--viramune--nvp-. Accessed August 28, 2017.

Uthman OA, Nachega JB, Anderson J, et al. Timing of initiation of antiretroviral therapy and adverse pregnancy outcomes: a systematic review and meta-analysis. *Lancet HIV* 2017;**4**(1):e21-e30.

VanderWeele TJ, Mukherjee B, Chen J. Sensitivity analysis for interactions under unmeasured confounding. *Stat Med.* 2012;31:2552-2564.

Vanderweele T. An Introduction to Interaction Analysis. In: Vanderweele T. *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York, NY: Oxford University Press; 2015:249-285.

Vannappagari V, Koram N, Albano J, et al. Association between in utero zidovudine exposure and nondefect adverse birth outcomes: analysis of prospectively collected data from the Antiretroviral Pregnancy Registry. *BJOG.* 2016;**123**:910-916.

Vart P, Nigatu YT, Jaglan A, et al. Joint effect of hypertension and elevated serum phosphorus on the risk of mortality in national health and nutrition examination survey-III. *J Am Heart Assoc.* 2015;4(5):pii:e001706.

Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol*. 1986;123(1):174-184.

Wang Q, Young J, Bernasconi E, et al. The prevalence of erectile dysfunction and its association with antiretroviral therapy in HIV-infected men: the Swiss HIV Cohort Study. *Antiviral Therapy* 2013; **18**: 337-344.

Wang Y, Wactawski-Wende J, Sucheston-Cambell LE, et al. The influence of genetic susceptibility and calcium plus vitamin D supplementation on fracture risk. *Am J Clin Nutr.* 2017;105(4):970-979.

Watts DH, Williams PL, Kacanek D, et al. Combination antiretroviral use and preterm birth. *J Infect Dis.* 2013;**207**(4):612-621.

Wensing AMJ, van Maarseveen NM, Nijhuis M. Fifteen years of HIV protease inhibitors: raising the barrier to resistance. *Antiviral Res* 2010; **85**: 59-74.

White AJ, DeRoo LA, Weinberg CR, et al. Binge drinking modifies the association between lifetime alcohol intake and breast cancer risk in moderate drinkers. *Am J Epidemiol.* 2017;186(5):541-549.

White DL, Li D, Nurgalieva Z, et al. Genetic variants of glutathione S-transferase as possible risk factors for hepatocellular carcinoma: a HuGE systemative review and meta-analysis. *Am J Epidemiol.* 2008;167(4):377-389.

Williams PL, Hazra R, Van Dyke RB, et al. Antiretroviral exposure during pregnancy and adverse outcomes in HIV-exposed uninfected infants and children using a trigger-based design. *AIDS* 2016;**30**:133-144.

Witte JS, Greenland S, Haile RW, et al. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. *Stat Med* 1994;**5**:612-621.

Witte JS, Greenland S. Simulation study of hierarchical regression. *Stat Med* 1996;**15**: 1161-1170.

Witte JS, Greenland S, Kim LL, et al. Multilevel modeling in epidemiology with GLIMMIX. *Epidemiology* 2000;**11**:684-688.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Human Genet*. 2011(89):82-93.

Yelland LN, Salter AB, Ryan P. Performance of the modified Poisson regression approach for estimating relative risks from clustered prospective data. *Am J Epidemiol.* 2011;174(8):984-992.

Young J, Glass TR, Bernasconi E, et al. Hierarchical modeling gave plausible estimates of associations between metabolic syndrome and components of antiretroviral therapy. *J Clin Epi* 2009;**62**:632-641.

Young J, Mucsi I, Rollet-Kurhajec KC, et al. Fibroblast growth factor 23: associations with antiretroviral therapy in patients co-infected with HIV and hepatitis C. *HIV Med* 2016;**17**:373-379.

Yu B, Wang Z. Estimating relative risks from common outcome using PROC NLP. *Comput Methods Programs Biomed*. 2008;90(2):179-186.

Zash RM, Williams PL, Sibiude J, et al. Surveillance monitoring for safety of in utero antiretroviral therapy exposures: current strategies and challenges. *Expert Opin Drug Saf* 2016; **15**:1501-1513.

Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics.* 1988;44(4):1049-1060.

Zhang J, Zhu J, Ding S, et al. Analysis of the synergistic effects of fasting plasma glucose and hypertension on cardiovascular autonomic neuropathy. *Cardiology*. 2015;132(1):58-64.

Zhang Q. Associating rare genetic variants with human diseases. Editorial. *Frontiers in Genetics.* 2015 (6): 133.

Zou GY. On the estimation of additive interaction by use of the four-by-two table and beyond. *Am J Epidemiol*. 2008;168(2):212-224.

Zou GY, Donner A. Extension of the modified regression model to prospective studies with correlated binary data. *Stat Methods Med Res.* 2013;22(6):661-670.