Title
Bayesian Meta-analysis Methods for Improving Accuracy and Adjusting for Publication Bias

Author
Gibson, Thomas

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Bayesian Meta-analysis Methods for Improving Accuracy and Adjusting for Publication Bias

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Thomas Gibson

2022

ABSTRACT OF THE DISSERTATION

Bayesian Meta-analysis Methods
for Improving Accuracy and
Adjusting for Publication Bias

by

Thomas Gibson
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2022
Professor Robert E. Weiss, Chair

Meta-analysis (MA) combines multiple studies to estimate a quantity of interest. Some existing MA models have shortcomings in the form of 1) inappropriate inference targets, 2) strong assumptions about how studies are sampled, and 3) prior distributions for variance parameters with inadequate shrinkage.

In Chapter 2 we build a three-random effect (3RE) Bayesian random effects MA model for observational contingency table data as an extension of the standard two-random effect (2RE) model. We add a random effect for the log-odds of having a risk factor with random effects for the log-odds of an event and for the log-odds ratio of the event for those with or without the risk factor. The 3RE model allows for calculation of more statistics than the 2RE model, and we define a novel estimand for statistics calculated from contingency tables – the expected value of a statistic

for a new study given the hyperparameters. The new estimand shows less bias and higher 95% credible interval coverage as compared with a naive plug-in estimator. We apply the model to a dataset of studies on patients presenting to the emergency department with syncope.

In Chapter 3 we propose a new approach to combining multiple selection models for publication bias using Bayesian stacking of posterior distributions. We demonstrate the effectiveness of stacking selection models through simulations and real datasets that exhibit symptoms of publication bias.

Chapter 4 proposes a new class of prior distributions for the covariance matrix of random effects in MA. The new priors allow random effects variances to be shrunk towards zero and for shrinkage of correlations between random effects. We show through both synthetic and real data examples that the new prior distributions lead to less diffuse posterior distributions and shorter 95% credible intervals in a 3RE MA model for observational data and an arm-based network meta-analysis (AB-NMA) model for randomized controlled trials.

The dissertation of Thomas Gibson is approved.

Onyebuchi Aniweta Arah

Donatello Telesca

Sudipto Banerjee

Robert E. Weiss, Committee Chair

University of California, Los Angeles

2022

To Amy and John,

for 30 years of unwavering support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Research reported in this publication was partially supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (NIH) under award R01HL111033.

VITA

2014	B.A. (Mathematics-Economics), Whitman College, Walla Walla, Washington.
2014	Summer Intern, Summer Institute for Training in Biostatistics, Emory University.
2016	Teaching Assistant, Department of Biostatistics, UCLA.
2016-2018	Graduate Student Researcher, Department of Biostatistics, UCLA. Worked with Professor Rob Weiss on a project involving emergency department patients presenting with syncope.
2017	M.S. (Biostatistics), UCLA, Los Angeles, California.
2018-2020	Teaching Assistant, Department of Biostatistics, UCLA.
2020	Graduate Student Researcher, Amgen Inc., Thousand Oaks, CA. Worked with the Data Science team at Amgen as a researcher in partnership with UCLA Department of Biostatistics.
2020-2021	Graduate Student Researcher, Department of Biostatistics, UCLA. Did research on forecasting the prevalence and incidence of Alzheimer's Disease under the direction of Professor Ron Brookmeyer.

PUBLICATIONS

Probst, M. A., Gibson, T., Weiss, R. E., Yagapen, A. N., Malveau, S. E., Adler, D. H., ... & Sun, B. C. (2020). Risk stratification of Older Adults who Present to the Emergency Department with Syncope: the FAINT score. Annals of Emergency Medicine, 75(2), 147-158.

Clark, C. L., Gibson, T. A., Weiss, R. E., Yagapen, A. N., Malveau, S. E., Adler, D. H., ... & Hollander, J. E. (2019). Do High-sensitivity Troponin and Natriuretic Peptide Predict Death or Serious Cardiac Outcomes After Syncope? Academic Emergency Medicine, 26(5), 528- 538.

Gibson, T. A., Weiss, R. E., & Sun, B. C. (2018). Predictors of short-term outcomes after syncope: a systematic review and meta-analysis. Western Journal of Emergency Medicine, 19(3), 517.

Holden, T. R., Shah, M. N., Gibson, T. A., Weiss, R. E., Yagapen, A. N., Malveau, S. E., ... & Sun, B. C. (2018). Outcomes of Patients with Syncope and Suspected Dementia. Academic Emergency Medicine, 25(8), 880-890.

Probst, M. A., Gibson, T. A., Weiss, R. E., Yagapen, A. N., Malveau, S. E., Adler, D. H., ... & Sun, B. C. (2018). Predictors of Clinically Significant Echocardiography Findings in Older Adults with Syncope: A Secondary Analysis. Journal of Hospital Medicine, 13(12), 823.

CHAPTER 1

Introduction

Medical practitioners must remain up to date on the best available scientific evidence to guide their decision-making while providing care for patients. In any given healthcare setting there are often multiple studies that have investigated a particular question with varying results. For example, there may be multiple studies examining the effect of a new drug compared to the current best treatment, and there is interest in combining the results from each study. *Meta-analysis* (MA) is a set of statistical models used for evidence synthesis, where results from multiple studies investigating the same question can be modeled together to get a pooled estimate of a quantity of interest. In this chapter we describe three current issues in meta-analysis and give a brief description of the solution we offer in each case.

Chapter 2 A common type of data to arise in medical studies is observational 2×2 contingency table data, where rows of the table are defined by the presence or absence of a risk factor (RF) and columns are defined by presence or absence of an adverse event. The data is observational because neither the number of subjects with/without the risk factor (row totals) or the number of subjects with/without the event (column totals) are fixed by investigators. The inference targets are *contingency table statistics* (CTSs), which measure the diagnostic utility of the RF. Some

commonly measured CTSs are sensitivity and specificity, which are the probability of having or not having the RF given the presence or absence of the event, respectively, and positive and negative predictive values, which are the probability of having or not having the event given the presence or absence of the RF, respectively.

The standard Bayesian meta-analysis model for 2×2 contingency table data is the random effects (RE) model, where we assume that the true underlying parameters describing each study population, such as the log-odds ratio, are different for each study and vary around some unknown *global* mean with some unknown RE variance. The global mean parameters and RE variances are *hyperparameters* in the model. The inference targets in a MA of observational 2×2 data are global CTSs. Study-specific CTSs can be calculated using study-specific parameters. Existing models for observational 2×2 data calculate global estimates for CTSs by plugging in global mean parameters for study-specific parameters and ignore the RE variance between studies; we call this the *plug-in estimator*. In Chapter 2 we define a novel estimand for CTSs, the expected value of a given CTS for a new study given all hyperparameters, which takes into account the RE variances. We propose a nested Monte Carlo procedure to sample from the posterior distribution of the new estimand, and we compare the new estimand to the naive plug-in estimator in a simulation study and analyze a set of real studies on patients presenting to the emergency department with syncope and assess the diagnostic utility of various regularly-measured covariates.

Chapter 3 Publication bias (PB) is a major threat to the validity of any meta-analysis. PB is an amalgamation of multiple sources of bias including language bias (favoring studies in English), familiarity bias (favoring studies from the lead investigator's own discipline), availability and cost bias (favoring studies that are free

and/or easily available), and reporting bias (unfavorable results within studies are not reported). PB may lead to an unrepresentative sample of studies and therefore biased meta-analysis results. In Chapter 3 we offer a new method of adjusting for PB. There is a diverse set of tools available to analysts who suspect their sample of studies may exhibit symptoms of PB. There are hypothesis tests that measure the asymmetry of a *funnel plot*, which plots estimated effect sizes against their standard errors, and return a p-value indicating the probability of observing as bad or worse asymmetry from chance alone. There are also sensitivity analyses, which assume varying degrees of publication bias and calculate bias-adjusted estimates under each scenario. If results change a lot with mild assumed bias, then results are sensitive to PB; if results remain fairly consistent even with severe assumed bias, the results are robust to PB. *Selection models* are a class of statistical models which define a mechanism through which studies are chosen to be included in the meta-analysis, and allow for bias-adjusted estimates of quantities of interest. However, posterior distributions for the bias-adjusted mean effect size can vary widely based on the selection model. For example, an assumption that selection of studies is based on one-sided p-values will tend to adjust the mean estimate towards the null more than an assumption that selection is based on two-sided p-values. Thus, a main issue is that results of the MA depend on the selection mechanism chosen by the meta-analyst. Recent approaches have used Bayesian model averaging (BMA) over multiple candidate models to increase robustness (Guan and Vandekerckhove, 2016; Maier et al., 2022). BMA performs well when the true model is one of the candidate models. However, a newer method of model combination called *Bayesian stacking* (Yao et al., 2018) has been shown to outperform BMA in situations where the true model is not in the list of candidate models. In Chapter 3 we argue that in the case

of publication bias, no model is "true" because the mechanisms of publication bias are too complex to capture in any single model. Therefore, we propose the stacking of multiple selection models for publication bias as a new robust method of adjusting for publication bias in meta-analysis.

Chapter 4 Many MA models have multiple random effects (REs) which are allowed to vary across studies. In a Bayesian RE model, one needs to model the covariance matrix associated with REs with an appropriate prior distribution. Covariance modeling in meta-analysis and network meta-analysis (NMA) has not been thoroughly researched. There are only a handful of prior distributions that have been discussed in the literature, and commonly used default prior distributions give lower prior density to some plausible values of variance parameters. For example, if we believe the variance of a certain random effect may be zero, the priors currently used in meta-analysis and NMA do not adequately allow for shrinkage of RE variance and will have posteriors that support larger variance values and not values near zero. Additionally, when there are few studies and/or few subjects per study, there is very little information in the data on correlations between REs when one of the REs has very small variance, and default covariance or correlation prior distributions will yield diffuse posterior distributions for correlation parameters. Inflated posterior variances and diffuse posteriors for correlations will in turn yield diffuse posterior distributions for other quantities of interest, such as absolute risks. In Chapter 4 we define a new class of prior distributions for the RE covariance matrix that allows for variances to shrink towards zero, and offers the option of shrinking correlations towards zero for REs that have very small variance. The new class of priors tend to yield more conservative mean estimates for quantities of interest, as

well as shorter 95% credible intervals, which we show through both synthetic and real data examples.

Each of Chapters 2, 3, and 4 is a standalone paper that can be read independently, and each Chapter has its own notation and data structure. Both Greek and non-Greek notation has a different meaning in each Chapter.

CHAPTER 2

Bayesian Meta-analysis of Observational Contingency Table Data with a Nested Monte Carlo Procedure for Estimating Global Effects

2.1 Introduction

Data from medical studies can often be tabulated in a 2×2 contingency table. The tables have columns stratified by a dichotomous outcome and rows stratified by a dichotomous covariate. Summary statistics from a 2×2 contingency table include positive/negative predictive value (PPV/NPV), sensitivity and specificity (Sens and Spec), and positive and negative likelihood ratios (LR+ and LR-), among others. We refer to a statistic that can be calculated as functions of some or all of the four values in a 2×2 contingency table as a *contingency table statistic* (CTS). For an individual study's table, this would mean using the counts in each cell to calculate an observed CTS, and for a population it would mean using the underlying multinomial cell probabilities to calculate a population CTS. CTSs describe the relationship between the outcome and the covariate, and calculating most CTSs requires conditioning on either rows or columns. Meta-analysis methods for contingency table data reflect this conditioning, and can generally be segregated into two groups that allow inference

for different CTSs.

Meta-analysis models for randomized controlled trials (RCTs) allow inference for *treatment* CTSs (T-CTSs), and naturally condition on the dichotomous covariate treatment/placebo. T-CTSs include the odds ratio (OR), relative risk (RR), risk difference (RD), and positive/negative predictive values (PPV/NPV). A standard random effects model for T-CTSs is given in Smith et al. (1995), with the log-odds ratio log(OR) as the main inference target. The model of Smith et al. (1995) has been extended to network meta-analysis with K treatment groups (Lu and Ades, 2004; Dias et al., 2013; Zhang et al., 2014). Models for diagnostic tests allow inference for *diagnostic* CTSs (D-CTSs) that condition on the presence or absence of an adverse event, denoted by E or $\overline{\text{E}}$. D-CTSs include sensitivity (Sens), specificity (Spec), ORs, and positive/negative likelihood ratios (LR+/LR-). Ma et al. (2016) reviews meta-analysis models that condition on event status, including the summary receiver operating characteristic (SROC) curve (Rutter and Gatsonis, 2001; Moses et al., 1993; Lian et al., 2019), bivariate random effects models (Reitsma et al., 2005; Chu and Cole, 2006; Arends et al., 2008; Chu et al., 2012; Guo et al., 2017; Hoyer and Kuss, 2018), and trivariate random effects models (Chu et al., 2009; Ma et al., 2018; Wynants et al., 2018). Models for T-CTSs and D-CTSs are similar in that they use binomial likelihoods conditioning on rows or columns, respectively. Multiple models (Chu et al., 2009; Rutter and Gatsonis, 2001; Ma et al., 2018) aim to estimate *global* statistics, synonymously referred to as "overall", "summary", or "population" statistics that are not study-specific.

An area of medical literature particularly suited to generating 2×2 data is emergency department (ED) visits for syncope (fainting), where around 5-10% of older syncope patients experience an adverse event in the 30 days after their initial ED

visit (Gibson et al., 2018). Many studies provide 2×2 tables of counts for dichotomous *risk factors* (RFs) that are regularly collected during an ED visit for syncope patients, including demographics, comorbidities, symptoms, and test characteristics. The syncope data is unique in that 1) it is *observational*, with neither row totals nor column totals fixed by study investigators, and for which we are interested in both T-CTSs and D-CTSs, and 2) given the large number (> 30) of regularly-measured covariates in the ED for syncope patients, we would like to "weed out" those covariates that are unrelated to the probability of 30-day adverse events.

To make inference on T- and D-CTSs together we propose a novel 3 random effect (3RE) Bayesian meta-analysis model as an extension of the model in Smith et al. (1995), with study-level random effects on the average log-odds of an event, the log(OR) of the event, and the log-odds of having the risk factor. Existing 3RE models (Chu et al., 2009; Ma et al., 2018; Wynants et al., 2018) model the probability of a positive diagnostic test simultaneously with sensitivity and specificity, but their methods use *plug-in estimators*, plugging in hyperparameters to calculate global statistics, to provide *median* estimates of global T-CTSs and D-CTSs. We instead posit that global *mean* effects are more desirable and define a novel estimand, the *expected value of a given statistic for a new study*, which accounts for all appropriate variability, and we outline a procedure to sample from the posterior distribution of the estimand. We use a fully Bayesian approach which has advantages in interpretation and flexibility, and our parameterization is different in that it builds off of Smith et al. (1995) with the log(OR) as the natural parameter. Whether or not a risk factor is related to 30-day adverse events corresponds to a natural scientific question in random effects meta-analysis of whether or not the mean log(OR) for a given risk factor is different from zero (Higgins et al., 2009). We introduce a mixture spike-and-

slab prior distribution (George and McCulloch, 1993, 1997; Kuo and Mallick, 1998; Ishwaran et al., 2005) on the mean parameter for the random effects on the log(OR) in the 3RE model, which allows us to calculate the posterior probability that the null hypothesis is true. The mixture prior places point mass on the probability that the mean log(OR) = 0 (the spike), and if not 0, models uncertainty in the mean log(OR) with a continuous prior distribution (the slab).

We present the 3RE meta-analysis model, a nested Monte Carlo procedure for calculating global CTSs, and a spike-and-slab prior on the global log(OR) in Section 2.2. Section 2.3 presents a simulation to show how well the model identifies true zero and non-zero effects and how accurately and precisely the nested Monte Carlo procedure estimates common CTSs as compared with plug-in estimators. Section 2.4 applies the model to our motivating syncope data. The paper closes with discussion.

2.2 Three RE meta-analysis model

In the usual meta-analysis, each study $i = 1, \ldots, S$ reports a 2×2 table of counts n_{ijk} with rows $j = 0, 1$ defined by the absence or presence of a risk factor (RF), denoted $\overline{\text{RF}}$ and RF, and columns $k = 0, 1$ defined by no adverse event ($\overline{\text{E}}$) or adverse event (E). Let $n_{i1} = n_{i10} + n_{i11}$ and $n_{i0} = n_{i00} + n_{i01}$ be the number of people with or without the risk factor, respectively, $N_i = n_{i1} + n_{i0}$ be the total sample size in study i, and π_{ij} is the probability of an adverse event for a patient in study i, group j as illustrated in Table 2.1. Assuming binomial sampling, the standard Bayesian random

	No Event	Event	Total
j = 0, RF Absent	n_{i00}	n_{i01}	n_{i0}
j = 1, RF Present	n_{i10}	n_{i11}	n_{i1}
j = 0, RF Absent	$1 - \pi_{i0}$	π_{i0}	1
j = 1, RF Present	$1 - \pi_{i1}$	π_{i1}	1

Table 2.1: Sample contingency table for study i with subject counts (top) and a probability representation conditional on presence or absence of the risk factor (bottom).

effects meta-analysis model is

$$n_{ij1}|\pi_{ij} \sim \text{Bin}(n_{ij}, \pi_{ij}) \tag{2.1}$$

$$\text{logit}(\pi_{ij}) = \begin{cases} \beta_i - \frac{\delta_i}{2} & j = 0 \\ \beta_i + \frac{\delta_i}{2} & j = 1, \end{cases} \tag{2.2}$$

where $\text{logit}(a) = \log(a/(1-a))$, $0 < a < 1$, β_i is a random intercept term for the log-odds of the event for study i and δ_i is a random effect for the log(OR) of the event in study i. We model δ_i and β_i as normal with unknown mean and variance

$$\beta_i|\beta_0, \sigma_\beta^2 \sim \text{N}(\beta_0, \sigma_\beta^2), \tag{2.3}$$

$$\delta_i|\delta_0, \sigma_\delta^2 \sim \text{N}(\delta_0, \sigma_\delta^2), \tag{2.4}$$

where population means β_0 and δ_0 and variances σ_β^2 and σ_δ^2 have priors $p(\beta_0)$, $p(\delta_0)$, $p(\sigma_\beta)$, and $p(\sigma_\delta)$ which we discuss in Section 2.2.3.

With this model we can make inference on T-CTSs. For observational data where we also want to make inference on D-CTSs, we expand model ((2.1)) - ((2.4)) to include a random effect $\psi_i = \text{P(RF in study } i)$, the probability of a subject having

the risk factor in study i. Assuming binomial sampling of n_{i1} from N_i as $\text{Bin}(N_i, \psi_i)$, we model $\nu_i = \text{logit}(\psi_i)$ as normal with unknown mean and cross-study variance σ_ν^2 on the logit scale

$$n_{i1}|\psi_i \sim \text{Bin}(N_i, \psi_i) \tag{2.5}$$

$$\nu_i|\nu_0, \sigma_\nu^2 \sim \text{N}(\nu_0, \sigma_\nu^2), \tag{2.6}$$

where the unknown population parameters ν_0 and σ_ν^2 have priors $f(\nu_0)$ and $f(\sigma_\nu^2)$.

2.2.1 Predictive contingency table statistics

There are study-specific and *global* versions of each CTS. Let $\boldsymbol{Y}$ be the data from all S studies, define $\boldsymbol{\theta}_i = (\beta_i, \delta_i, \nu_i)'$, the parameter vector for the i^{th} study, and let $\boldsymbol{\gamma} = (\beta_0, \sigma_\beta, \delta_0, \sigma_\delta, \nu_0, \sigma_\nu)'$ be the vector of hyperparameters. The unknown study-specific CTS_i's are functions of

$$\begin{aligned} \pi_{i1} &= \text{expit}(\beta_i + \delta_i/2) \\ \pi_{i0} &= \text{expit}(\beta_i - \delta_i/2) \\ \psi_i &= \text{expit}(\nu_i), \end{aligned} \tag{2.7}$$

where $\text{expit}(x) = 1/(1 + \exp(-x))$. The T-CTSs PPV, NPV, RD, and RR for study i are

$$\begin{aligned} \text{PPV}_i &= \text{P}(\text{E}|\text{RF}) = \pi_{i1} & \text{NPV}_i &= \text{P}(\overline{\text{E}}|\overline{\text{RF}}) = 1 - \pi_{i0} \\ \text{RR}_i &= \frac{\text{P}(\text{E}|\text{RF})}{\text{P}(\text{E}|\overline{\text{RF}})} = \frac{\pi_{i1}}{\pi_{i0}} & \text{RD}_i &= \text{P}(\text{E}|\text{RF}) - \text{P}(\text{E}|\overline{\text{RF}}) = \pi_{i1} - \pi_{i0}, \end{aligned} \tag{2.8}$$

and the D-CTSs Sens, Spec, LR+, and LR− are

$$\begin{aligned} \text{Sens}_i &= \text{P}(\text{RF}|\text{E}) = \frac{\pi_{i1}\psi_i}{\pi_{i1}\psi_i + \pi_{i0}(1-\psi_i)} & \text{LR}-_i &= \frac{1-\text{Sens}_i}{\text{Spec}_i} \\ \text{Spec}_i &= \text{P}(\overline{\text{RF}}|\overline{\text{E}}) = \frac{(1-\pi_{i0})(1-\psi_i)}{(1-\pi_{i0})(1-\psi_i) + (1-\pi_{i1})\psi_i} & \text{LR}+_i &= \frac{\text{Sens}_i}{1-\text{Spec}_i}. \end{aligned} \tag{2.9}$$

Each CTS_i is then a function $g(\boldsymbol{\theta}_i)$ of the study-specific parameters $\boldsymbol{\theta}_i$ for appropriate choice of $g(\cdot)$.

Usually the purpose of meta-analysis is to consolidate information from multiple studies, and we are interested in *global* rather than study-specific CTSs. Existing methods use *plug-in estimators*, $\text{CTS}_{\text{plug}}(\beta_0, \delta_0, \nu_0)$ as estimates of global CTSs, by plugging in

$$\begin{aligned} \pi_1 &= \text{expit}(\beta_0 + \delta_0/2) \\ \pi_0 &= \text{expit}(\beta_0 - \delta_0/2). \\ \psi &= \text{expit}(\nu_0), \end{aligned}$$

in equations ((2.8)) - ((2.9)) in place of study-specific π_{i1}, π_{i0}, and ψ_i. The plug-in method ignores 1) the nonlinear relationship between the mean hyperparameters $(\beta_0, \delta_0, \nu_0)$ and the global CTSs, and 2) the heterogeneity present across studies represented by σ_β, σ_δ and σ_ν. In contrast, we define the target estimand as the *predictive mean* $\text{CTS}_0(\boldsymbol{\gamma})$ *of a CTS for a new study* given $\boldsymbol{\gamma}$, $\text{CTS}_0(\boldsymbol{\gamma}) = \text{E}[g(\boldsymbol{\theta}_{S+1})|\boldsymbol{\gamma}]$, where $\boldsymbol{\theta}_{S+1} = (\beta_{S+1}, \delta_{S+1}, \nu_{S+1})$ and $\beta_{S+1}, \delta_{S+1}$, and ν_{S+1} are distributed as in ((2.3)), ((2.4)), and ((2.6)), and study index i can take on value $S+1$. For brevity, define $\text{CTS}_0 \equiv \text{CTS}_0(\boldsymbol{\gamma})$ and $\text{CTS}_{\text{plug}} \equiv \text{CTS}_{\text{plug}}(\beta_0, \delta_0, \nu_0)$.

For given $\boldsymbol{\gamma}$, we approximate $\text{CTS}_0 = \text{E}[g(\boldsymbol{\theta}_{S+1})|\boldsymbol{\gamma}]$ with a Monte Carlo estimate

$$\begin{aligned} \text{E}[g(\boldsymbol{\theta}_{S+1})|\boldsymbol{\gamma}] &= \int g(\boldsymbol{\theta}_{S+1}) P(\boldsymbol{\theta}_{S+1}|\boldsymbol{\gamma}) d\boldsymbol{\theta}_{S+1} \\ &\approx \frac{1}{L} \sum_{l=1}^{L} g(\boldsymbol{\theta}_{S+1}^{(l)}) \end{aligned} \tag{2.10}$$

where integer L is chosen to make Monte Carlo error in (2.10) desirably small, and $\boldsymbol{\theta}_{S+1}^{(l)}$ are drawn from the predictive distribution $P(\boldsymbol{\theta}_{S+1}|\boldsymbol{\gamma})$. To draw samples $m = 1, \ldots, M$ from the posterior distribution $P(\text{CTS}_0|\boldsymbol{Y})$ within a Markov chain Monte Carlo (MCMC) algorithm we approximate the integral $\int g(\boldsymbol{\theta}_{S+1}) P(\boldsymbol{\theta}_{S+1}|\boldsymbol{\gamma}) d\boldsymbol{\theta}_{S+1}$ in each iteration m with a Monte Carlo calculation. Given M MCMC samples $\boldsymbol{\gamma}^{(m)}$, $m = 1, \ldots, M$ from the posterior of $P(\boldsymbol{\gamma}|\boldsymbol{Y})$, for each m we

1. Take L draws $\boldsymbol{\theta}_{S+1}^{(m,l)}$, $l = 1, \ldots, L$ from the predictive distribution $P(\boldsymbol{\theta}_{S+1}|\boldsymbol{\gamma}^{(m)})$ and calculate the CTS $g(\boldsymbol{\theta}_{S+1}^{(m,l)})$ for each of the L draws,

2. Estimate $\text{CTS}_0(\boldsymbol{\gamma}^{(m)}) = \text{E}[g(\boldsymbol{\theta}_{S+1})|\boldsymbol{\gamma}^{(m)}] \approx \frac{1}{L} \sum_{l=1}^{L} g(\boldsymbol{\theta}_{S+1}^{(m,l)})$.

Sampling $\boldsymbol{\gamma}^{(m)}$ and $\text{CTS}_0(\boldsymbol{\gamma}^{(m)})$ in each iteration of MCMC sampling yields approximate samples from the posterior distribution for the expectation of the CTS given data $\boldsymbol{Y}$ and $\boldsymbol{\gamma}$, $p(\text{CTS}_0|\boldsymbol{Y})$, where uncertainty in CTS_0 is due to uncertainty in the parameters $\boldsymbol{\gamma}$. We refer to this method as the MC procedure.

2.2.2 Spike-and-slab prior for the log-odds ratio

We want to formally test the null hypothesis $H_0\colon \delta_0 = 0$ that the log(OR) is 0 against $H_A\colon \delta_0 \neq 0$. The Bayesian approach to testing builds an encompassing model where both H_0 and H_A have positive probability, for example, with a spike-and-slab (SAS)

prior $p_1(\delta_0)$ for δ_0

$$\delta_0 = \begin{cases} \delta & \rho = 1 \\ 0 & \rho = 0 \end{cases} \tag{2.11}$$

$$\rho \sim \text{Bernoulli}(p) \tag{2.12}$$

$$\delta \sim \text{N}(0, b_\delta^2), \tag{2.13}$$

where p is the prior probability that $\delta_0 = 0$. In the absence of other prior information we usually choose $p = 0.5$. We call the 3RE model ((2.1)) - ((2.6)) with prior ((2.11)) - ((2.13)) the 3RE-SAS model. In the absence of prior information one can set prior standard deviation $b_\delta = 2$ to give support to values of $\delta_0 \in (-4, 4)$, where a log(OR) δ of -4 or 4 corresponds to an OR of 0.02 or 55. A special case has $\rho = 1$, and $\delta_0 \equiv \delta$, which is a continuous prior for δ_0.

2.2.3 Prior distributions

For the mean parameters β_0 and ν_0 we propose normal distributions with known means a_β, a_ν and variances b_β^2, b_ν^2

$$\beta_0 \sim \text{N}(a_\beta, b_\beta^2) \tag{2.14}$$

$$\nu_0 \sim \text{N}(a_\nu, b_\nu^2). \tag{2.15}$$

Prior means a_β and a_ν are prior guesses at the mean log-odds of an event and log-odds of the risk factor, respectively, and the standard deviations b_β and b_ν are chosen to be large enough to give support to all plausible values of the parameters. As a default we assign each of the prior standard deviations σ_β, σ_δ, and σ_ν weakly informative

half-Cauchy prior distributions truncated above 5

$$\sigma_\beta \sim \text{half-Cauchy}(A_\beta)\mathbb{1}_{[\sigma_\beta<5]} \tag{2.16}$$

$$\sigma_\delta \sim \text{half-Cauchy}(A_\delta)\mathbb{1}_{[\sigma_\delta<5]} \tag{2.17}$$

$$\sigma_\nu \sim \text{half-Cauchy}(A_\nu)\mathbb{1}_{[\sigma_\nu<5]} \tag{2.18}$$

where $\sigma \sim$ half-Cauchy(A) with scale parameter $A > 0$ has density $p(\sigma) \propto (A^2 + \sigma^2)^{-1}\mathbb{1}_{[\sigma>0]}$ (Gelman, 2006). The scale parameters A_β, A_δ, and A_ν should generally be set between 0.25 and 1. We should expect standard deviations σ_β, σ_δ, and σ_ν to be below 1.5, as values above 1.5 may signal problems with model fit/appropriateness or the data because heterogeneity that large is unlikely to occur naturally. Taking $A = 0.25$ yields a prior probability $P(\sigma < 1.5) \approx 0.9$, while $A = 1$ yields $P(\sigma < 1.5) \approx 0.65$. The choice of A matters less with more studies in the meta-analysis. With fewer than 10 studies we recommend $A \in (0.25, 0.5)$. With very few studies (2 or 3), large standard deviations can lead to specificity having posterior mass very close to 1, inducing calculation problems and occasional unrealistically large values for the statistic LR+ $=$ Sens$/(1-$Spec$)$. A priori, we do not believe that LR$+_0 > 30$ in our syncope data analysis. Thus it would be sensible to restrict LR$+_0 < 30$ in the prior, on top of the prior specification for $\boldsymbol{\gamma}$. In practice, the restriction LR$+_0 < 30$ or some other value may or may not be needed. The need might be indicated by a long right tail in the posterior for LR$+_0$, possibly indicated by a posterior SD of LR$+_0$ larger than the mean, or any posterior probability of LR$+_0 > 30$. If needed it can be implemented with a post-hoc removal of any MCMC posterior samples where LR$+_0 > 30$ or other chosen upper bound. In our simulations and data analysis we restrict LR$+_0 < 30$.

2.2.4 Special case: fixed effect for the log-odds ratio

If we have evidence that the standard deviation σ_δ of random effects δ_i is small, i.e. $\sigma_\delta \approx 0$, we can instead model a fixed effect for the log(OR) δ_0 where $\delta_1 = \cdots = \delta_S = \delta_0$. Modify equation ((2.2)) to

$$\text{logit}(\pi_{ij}) = \begin{cases} \beta_i - \frac{\delta_0}{2} & j = 0 \\ \beta_i + \frac{\delta_0}{2} & j = 1, \end{cases} \tag{2.19}$$

where δ_0 has the SAS prior ((2.11)) - ((2.13)). We call this the 2RE-SAS model. The posterior probability $P(\delta_0 = 0|Y)$ is then the probability that $\delta_i = 0$ for every study i in the analysis, as well as for a future study $S+1$. Because $\delta_{S+1} = 0$ implies $\pi_{[S+1]1} = \pi_{[S+1]0}$, then $\text{RD}_{[S+1]} = 0$, $\text{RR}_{[S+1]} = 1$, $\text{LR}+_{[S+1]} = 1$, and $\text{LR}-_{[S+1]} = 1$, and the posterior probability $P(\delta_0 = 0|Y)$ is also the posterior probability that $\text{RD}_0 = 0$, $\text{RR}_0 = 1$, $\text{LR}+_0 = 1$, and $\text{LR}-_0 = 1$, where 0 and 1 are the null values for the respective CTSs. This is in contrast to the 3RE-SAS model, where $\delta_0 = 0$ does not imply that RR, RD, LR+ and LR− are exactly equal to their null values, as $\sigma_\delta > 0$.

2.3 Simulation studies

We perform three simulations to

1. Determine appropriate choices for L in calculating CTS_0 in the MC procedure;
2. Compare the MC procedure CTS_0 to the plug-in estimator CTS_{plug} with known target values CTS_0; and

3. Assess the posterior probability of the null hypothesis $\delta_0 = 0$ in the 3RE-SAS model.

Each simulation varies different factors, and we refer to each combination of factors as a scenario.

Certain simulation parameters are held constant in all three simulations. We fix $\beta_0 = \nu_0 = \log(.15/.85)$ and independently draw the number of subjects for each study from a discrete Uniform$(250, 2500)$ distribution to match values typical of the syncope data analysis. We set 4000 MCMC iterations in each of 4 chains, discard the first 2000 iterations as burn-in and use a thin of 2, leaving 4000 MCMC samples from each posterior. We set prior means $a_\beta = a_\nu = -1$, prior variances $b_\beta = b_\delta = b_\nu = 4$, and $A_\beta = A_\delta = A_\nu = \frac{1}{\sqrt{2}} \approx .707$. For the SAS prior we use a prior probability $P(\delta_0 = 0) = 0.5$. Initial values for mean parameters are drawn independently from a Uniform(-1, 1) distribution, and initial values for random effect standard deviations σ_β, σ_δ, and σ_ν are drawn independently from a Uniform(0.2, 1) distribution. We fit all models using JAGS (Plummer et al., 2003) in R (R Core Team, 2021).

2.3.1 Simulation 1: Choice of L for Monte Carlo procedure

Simulation 1 consists of two smaller simulations. First we compare posterior standard deviations (SDs) and 95% credible interval (95% CI) lengths for LR$+_0$ for several choices of L with 95% CIs having 2.5% probability content in each tail. We take the number of studies $S = 10$, fix $\delta_0 = 2$, and generate $K = 100$ datasets with random effect standard deviations (RESDs) $\sigma_\beta = \sigma_\delta = \sigma_\nu = 0.5$ or $= 1.0$.

Table 2.2 shows average posterior standard deviations (SDs) and 95% CI lengths of LR$+_0$ for each L for the two simulation scenarios. As L increases $\in \{1, 10, 100,$

	$\sigma_\beta = \sigma_\delta = \sigma_\nu = 0.5$		$\sigma_\beta = \sigma_\delta = \sigma_\nu = 1$	
L	SD	95% Length	SD	95% Length
1	1.969	7.502	3.869	14.876
10	0.868	3.360	2.295	8.646
100	0.642	2.508	1.730	6.472
1000	0.615	2.400	1.640	6.138
10000	0.612	2.390	1.633	6.099

Table 2.2: Simulation 1: Average posterior standard deviation (SD) and 95% interval length of LR$+_0$ for different values of L across $K_{11} = 100$ simulation iterations. Second and third columns are results when data was generated with $\sigma_\beta = \sigma_\delta = \sigma_\nu = 0.5$, and in fourth and fifth columns $\sigma_\beta = \sigma_\delta = \sigma_\nu = 1$. SDs and 95% interval lengths decrease with increasing L with diminishing returns.

$1000, 10000\}$, Monte Carlo error from Equation (2.10) of $\text{CTS}_0(\boldsymbol{\gamma})$ decreases and this can be inferred from Table 2.2 because the average posterior SD and 95% CI length decrease to an apparent limit. For sufficiently large L the error is negligible in comparison to the posterior SD of $\text{CTS}_0(\boldsymbol{\gamma})$.

The effect of further increasing L has decreasing impact on SD and 95% CI length. If true RESDs are moderate (≈ 0.5), then $L = 100$ seems sufficient. However, if RESDs are large then $L = 1000$ seems preferable and picking a smaller L might have an impact on inferences. Taking $L = 10000$ offers little additional precision.

In a second simulation we measure uncertainty in approximating $\text{CTS}_0(\boldsymbol{\gamma})$ with equation (2.10) for RESDs = 0.5 or 1.0. We sample 10000 replicates of LR$+_0$. The sampling standard deviation of LR$+_0$ is 1.703 when the RESDs are 0.5 and is 4.674 when the RESDs are 1. Therefore the standard error $\text{SE}_{\text{within}}$ of the $\text{CTS}_0(\boldsymbol{\gamma})$ calculation in equation (2.2.1) for LR$+_0$, is $1.703/\sqrt{L}$ or $4.674/\sqrt{L}$, and the user must decide what value of L makes $\text{SE}_{\text{within}}$ desirably small. The bottom row of Table 2.2 with $L = 10000$ offers a close approximation to the *between-MC* standard

deviation $\text{SD}_{\text{between}}$ for CTS_0. The user's choice of L should depend on the ratio of within-MC SE to between-MC SD. We recommend at least $\text{SE}_{\text{within}}/\text{SD}_{\text{between}} < 0.1$. For Simulations 2 and 3 and the Syncope data analysis we use $L = 1000$.

2.3.2 Simulation 2: estimating CTS_0

Simulation 2 evaluates how accurately the posterior mean of CTS_0, $\text{E}[\text{CTS}_0|Y]$, from the 3RE model estimates the true CTS values $\text{LR}+_0$, $\text{LR}-_0$, PPV_0, NPV_0, Sens_0, and Spec_0 for a new study and how far off the naive plug-in estimator CTS_{plug} is from CTS_0. We fix $\delta_0 = 2$, vary $\sigma_\delta \in \{0.1, 0.25, 0.5\}$, and vary the number S of studies per meta analysis with $S \in \{10, 30, 50\}$ in a 3×3 factorial design for 9 scenarios.

Let T_k be the posterior mean of the estimand of interest in simulation iteration k, $k = 1, \ldots, K$, and define the simulation mean $\overline{T} = \frac{1}{K}\sum_{k=1}^{K} T_k$ and variance $V_T = \frac{1}{K-1}\sum_{k=1}^{K}(T_k - \overline{T})^2$, and let μ be the known target value that T_k is estimating. We choose K to make the Monte Carlo standard error (MCSE) of relative bias, $\text{MCSE(rBias)} = \sqrt{V_T/(K\mu^2)}$, sufficiently small for each scenario. In this simulation we define T_k as the posterior mean $T_k = \text{E}[\text{CTS}_0|Y_k]$ for the CTSs LR+, LR−, NPV, PPV, Sens, and Spec. We calculate target values μ for each scenario by generating 100,000 probability tables from equations ((2.2)) - ((2.7)) using the known hyperparameters $\boldsymbol{\gamma}$, calculating the desired CTSs for each probability table, and averaging each CTS across tables. Using a preliminary set of 100 simulations, we calculate $K_2 = 2500$ such that $\text{MCSE(rBias)} < 0.0025$ for all scenarios and CTSs.

We record the posterior mean, SD, and 95% CI for every CTS for both CTS_0 and CTS_{plug} and calculate bias, 95% CI coverage, average 95% CI length, and root mean-squared error (RMSE) for the CTSs LR+, LR−, NPV, PPV, Sens, and Spec.

Figure 2.1 shows bar plots of 100∗bias for each CTS, combination of S and σ_δ, and CTS_0 or CTS_{plug}. Dotted bars plot the bias for CTS_{plug} and solid bars plot the bias for CTS_0. For sample sizes $S = 30$ and 50, CTS_{plug} has bias that is both significantly different from zero and larger than bias for CTS_0, with CTS_{plug} bias surprisingly increasing as sample size increases for NPV and Spec, which we discuss further in Section 2.5. There is a similar pattern in Figure 2.2 for RMSE, where RMSE for CTS_0 is smaller than for CTS_{plug} when $S \in \{30, 50\}$, although not always significantly. Figure 2.3 shows that 95% CI coverage of CTS_0 is always greater than or equal to 95%, while CTS_{plug} 95% coverage falls significantly below 95% for NPV, PPV, and Spec as S increases for all values of σ_δ, and for LR+ when $\sigma_\delta = 0.5$. We expect CTS_0 to have larger uncertainty than CTS_{plug} because CTS_0 accounts for variation in random effects, so an unexpected result in Figure 2.4 was that 95% CIs using CTS_0 are shorter on average than those of CTS_{plug} for LR− and Sens.

Overall, CTS_0 from the MC procedure tends to have lower bias and lower RMSE than the plug-in estimator as S increases. It also maintains at least nominal coverage as S increases, while the plug-in estimator sees coverage probabilities fall below nominal levels for multiple CTSs.

2.3.3 Simulation 3: true zero and non-zero effects

For simulation 3, we vary $\delta_0 \in (0, 1, 2)$ and $\sigma_\delta \in (0.1, 0.25, 0.5)$ in a 3×3 factorial experiment with 9 scenarios. The number of simulation iterations K_3 is set equal to 1000 so that the posterior probability $T_k = P(\delta_0 = 0|Y_k)$ has $\text{MCSE}(\overline{T}) < 0.005$. For each scenario we generate $K_3 = 1000$ datasets, where every dataset Y_k has $S = 10$ studies. For each iteration $k = 1, \ldots, 1000$ we fit the 3RE-SAS model to data Y_k and

calculate $P(\delta_0 = 0|Y_k)$ as the proportion of MCMC samples in which $\rho = 0$.

		Posterior $P(\delta_0 = 0\|Y)$		
	σ_δ	$\delta_0 = 0$	$\delta_0 = 1$	$\delta_0 = 2$
Mean	0.10	0.9286	0.0001	0.0000
	0.25	0.8976	0.0011	0.0000
	0.50	0.8494	0.0249	0.0001
SD	0.10	0.0655	0.0003	0.0000
	0.25	0.0920	0.0068	0.0000
	0.50	0.1251	0.0566	0.0002

Table 2.3: Simulation mean and SD of posterior $P(\delta_0 = 0|Y)$ across 1000 simulation iterations for each combination of $\sigma_\delta \in \{0.1, 0.25, 0.5\}$ and $\delta_0 \in \{0, 1, 2\}$. The first three rows give the mean over simulations of the posterior probability $P(\delta_0 = 0|Y)$ that the logs odds ratio δ_0 is zero, and the last three rows give the SD over simulations of $P(\delta_0 = 0|Y)$.

Figure 2.5 presents boxplots of the distribution of $P(\delta_0 = 0|Y_k)$ for each combination of $(\delta_0, \sigma_\delta)$ and we report the mean and standard deviation of $P(\delta_0 = 0|Y_k)$ for each situation in Table 2.3. There is a clear distinction in Figure 2.5 between simulations with $\delta_0 = 0$ (left-most boxplot in each panel) and simulations with $\delta_0 \neq 0$ (middle and right-most boxplot in each panel). The simulation mean of $P(\delta_0 = 0|Y_k)$ is near zero for true non-zero effects $\delta_0 \in (1, 2)$, and ranges from 0.79 to 0.91 for true mean zero effects $\delta_0 = 0$.

2.4 Syncope data analysis: assessing diagnostic utility of regularly measured covariates

Syncope, defined as transient loss of consciousness with rapid and spontaneous recovery, accounts for approximately 1.3 million emergency department (ED) visits every year in the United States (Probst et al., 2015). Syncope is often harmless, but may

be a harbinger of an impending serious cardiac event. ED physicians have difficulty determining which patients are at high risk for an event, and as a result admit up to 85% of older adults presenting with syncope (Birnbaum et al., 2008) even though only 5-10% of those presenting to the ED will have an event in the ensuing 30 days (Gibson et al., 2018). Given the difficulty of predicting serious cardiac events, we wish to measure the diagnostic value of regularly measured risk factors and determine which risk factors may have zero diagnostic value in predicting 30-day adverse events. Potential risk factors include demographics/comorbidities, symptoms, physical findings, and biomarkers.

There are 12 studies which each report information on some but not all risk factors; we meta-analyze 31 risk factors for which at least 2 studies provided a 2×2 table. For each analysis we first assess $P(\delta_0 = 0|Y)$ using the 3RE-SAS model and then we re-run the model with a continuous prior for δ_0 and summarize posteriors of $\text{LR}+_0$, $\text{LR}-_0$, PPV_0, and NPV_0. We remove samples in which $\text{LR}+_0 > 30$; the largest fraction of samples removed was 1.35%. To account for the trimmed posterior sample size, in each analysis we set 5100 MCMC iterations in each of 4 chains, discard the first 1000 iterations as burn-in and use a thin of 2, leaving at least 8000 MCMC samples from each posterior after trimming. If the posterior probability of the risk factor having no effect is > 0.5, one would usually forego computing CTS_0 estimates because they would be close to their null values and would be unlikely to provide any diagnostic value.

Tables 2.4 and 2.5 detail results for the 20 RFs with $P(\delta_0 = 0|Y) < 0.5$, and Tables 2.6 and 2.7 detail results for the 11 RFs with $P(\delta_0 = 0|Y) > 0.5$. Rows for each Table are sorted by $P(\delta_0 = 0|Y)$, labeled "Spike", from lowest to highest. Columns of Tables 2.4 and 2.6 give posterior means and 95% CIs for the CTSs LR+,

LR−, NPV, PPV, and columns of Tables 2.5 and 2.7 give posterior means and CIs for Sens and Spec. Risk factors with more studies have smaller posterior SDs, and risk factors with high posterior probability of specificity near 1 tend to have wider CIs for $LR+_0$. In many circumstances, a threshold of $LR+ > 10$ or $LR- < 0.1$ are used to either rule in or rule out an impending adverse event with the presence or absence of a risk factor, respectively (Deeks and Altman, 2004; Ranganathan and Aggarwal, 2018). No variables have posterior mean $LR+_0 > 6.62$ or $LR-_0 < 0.44$, which highlights the difficulty physicians face in determining which syncope patients are at high risk of an adverse event. Results for all 31 RFs are available in Web Table 3 of the Supporting Information, which also includes results for sensitivity and specificity.

The biomarkers troponin, urea, and creatinine have the fewest studies in this meta-analysis, but appear to be promising diagnostic predictors of adverse events with large mean values of $LR+_0$. The RFs history CHF (congestive heart failure) and Dyspnea (shortness of breath) have the highest mean PPVs. Figure 2.6 shows contour plots of $LR+_0$ vs $LR-_0$ for the four risk factors with the smallest posterior $P(\delta_0 = 0|Y)$, which are age, male gender, CHF, and history of heart disease. We see varying degrees of correlation between LR+ and LR−, with a correlation of -0.15 for age and of -0.94 for male gender.

2.5 Discussion

Given information such as the probability of the risk factor ψ_{S+1}, in a future study $S+1$, one can incorporate the information to more accurately predict $\text{CTS}_{[S+1]}$'s. For example, say we know that 5%, 10%, or 25% of people in some new population have

elevated blood troponin. Table 2.8 lists posterior summaries for CTS_0s given known values of P(elevated troponin in new study) $= \psi_{S+1} \in \{.05, .10, .25\}$. We see that rising prevalence of elevated blood troponin corresponds with increasing sensitivity, and decreasing specificity, LR− and LR+, while there is no effect on NPV and PPV because NPV and PPV condition on the presence or absence of the risk factor.

The Monte Carlo-within MCMC procedure to calculate CTS_0 is conveniently implemented with a post-processing step after MCMC fitting, rather than implemented within MCMC. Given posterior draws $\boldsymbol{\gamma}^{(m)}, m = 1, \ldots, M$, we estimate $\text{CTS}_0(\boldsymbol{\gamma}^{(m)})$ for each m outside of the MCMC algorithm.

In simulation 2 comparing CTS_0 and the plug-in estimators, the bias for NPV_{plug} and $\text{Spec}_{\text{plug}}$ increased with the number of studies S. The CTSs, either plug-in or predictive are non-linear functions of the means or of the random effects. NPV and Spec in particular have upper bounds of 1 and their posteriors or predictive distributions tend to have long left tails. With increasing S, posteriors of the mean parameters β_0, δ_0, and ν_0 will become more bell curve shaped with smaller variances. In contrast, with increasing S, the random effects variances will converge to their true values and will not decrease. For NPV_{new} and Spec_{new}, these long left tails remain with increasing S, while for NPV_{plug} and $\text{Spec}_{\text{plug}}$, the decreasing variance with increasing S means that the non-linear transformation is better and better approximated by a linear transformation and the long left tails are reduced with increasing S. Thus the plug-in estimators' posterior means tend to look more like the non-linear function applied to the posterior means of β_0, δ_0, and ν_0 while posterior means for NPV_{new} and Spec_{new} will be consistently less than NPV_{plug} and $\text{Spec}_{\text{plug}}$.

One can incorporate information from studies that only report a log(OR) $\widehat{\delta}_i$ and

its standard error $\mathrm{SE}(\widehat{\delta}_i)$. We model $\widehat{\delta}_i$ as drawn from a normal distribution centered around its true log(OR) δ_i with variance equal to $\mathrm{SE}(\widehat{\delta}_i)^2$

$$\widehat{\delta}_i|\delta_i \sim \mathrm{N}(\delta_i, \mathrm{SE}(\widehat{\delta}_i)^2),$$

where the true log(OR) δ_i is distributed as in ((2.4)). Studies reporting pairs $\left(\widehat{\delta}_i, \mathrm{SE}(\delta_i)\right)$ provide information on the mean and variance $(\delta_0, \sigma_\delta^2)$ of the random effects for the log(OR). If there is a set of additional studies reporting 2×2 tables *with fixed row-totals*, we can model them using ((2.1)) - ((2.3)). They provide information on $(\delta_0, \sigma_\delta^2)$ and the mean and variance $(\beta_0, \sigma_\beta^2)$ of the random effects for the log-odds of the event but not on the value of ν_i. If the set of additional studies instead have *fixed column-totals*, one can reformulate the model using the "opposite" parameterization, where the parameters $(\beta_i, \delta_i, \nu_i)$ represent the log-odds of the risk factor, the diagnostic log(OR), and the probability of the event respectively in study i. Column totals are often fixed for case-control or diagnostic studies.

If there is suspected correlation between the random effects β_i, δ_i, and ν_i, they may be modeled with a multivariate normal distribution

$$\begin{pmatrix} \beta_i \\ \delta_i \\ \nu_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \beta_0 \\ \delta_0 \\ \nu_0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\beta^2 & \rho_{\beta\delta}\sigma_\beta\sigma_\delta & \rho_{\beta\nu}\sigma_\beta\sigma_\nu \\ \rho_{\beta\delta}\sigma_\beta\sigma_\delta & \sigma_\delta^2 & \rho_{\delta\nu}\sigma_\delta\sigma_\nu \\ \rho_{\beta\nu}\sigma_\beta\sigma_\nu & \rho_{\delta\nu}\sigma_\delta\sigma_\nu & \sigma_\nu^2 \end{pmatrix} \right)$$

with an inverse-Wishart prior on the covariance matrix. In this paper we model the random effects as independent because the syncope data analysis showed nearly identical results.

There is also potential for meta-analysis results to be used for prior specification

in the design and analysis of a future study. The posterior probability $P(\delta_0 = 0|Y)$ can be used to either screen out variables in future analyses, or can be used as prior probabilities in a Bayesian variable selection model using spike-and-slab priors for regression coefficients, as was done in Probst et al. (2020).

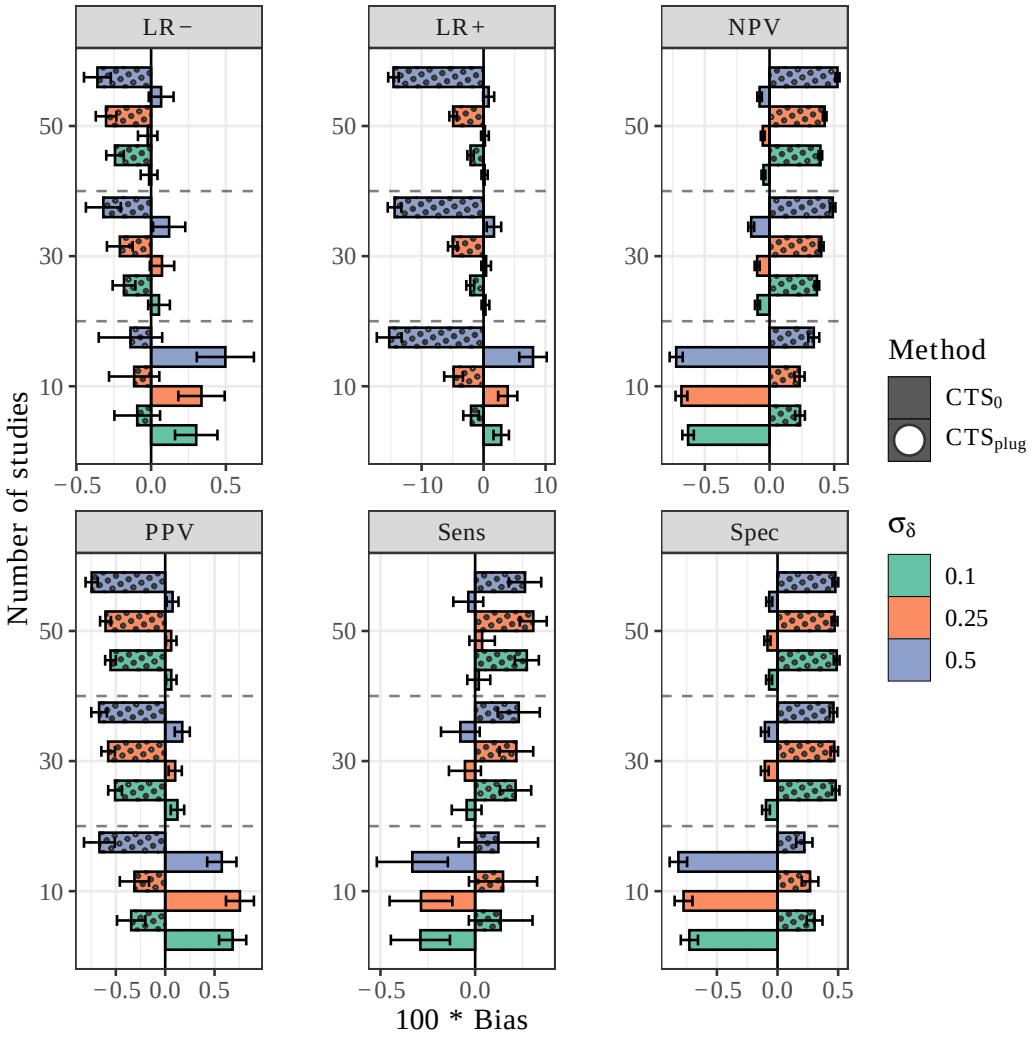

Figure 2.1: Bar plot of 100 $*$ bias for each CTS and for each combination of $S \in \{10, 30, 50\}$ on the y-axis and $\sigma_\delta \in \{0.1, 0.25, 0.5\}$ indicated by green, red, and blue, respectively. Dotted bars plot bias for the plug-in estimator CTS_{plug} and solid bars plot bias for the Monte Carlo estimator CTS_0. Error bars represent $\pm 1.96 \times$ MCSE.

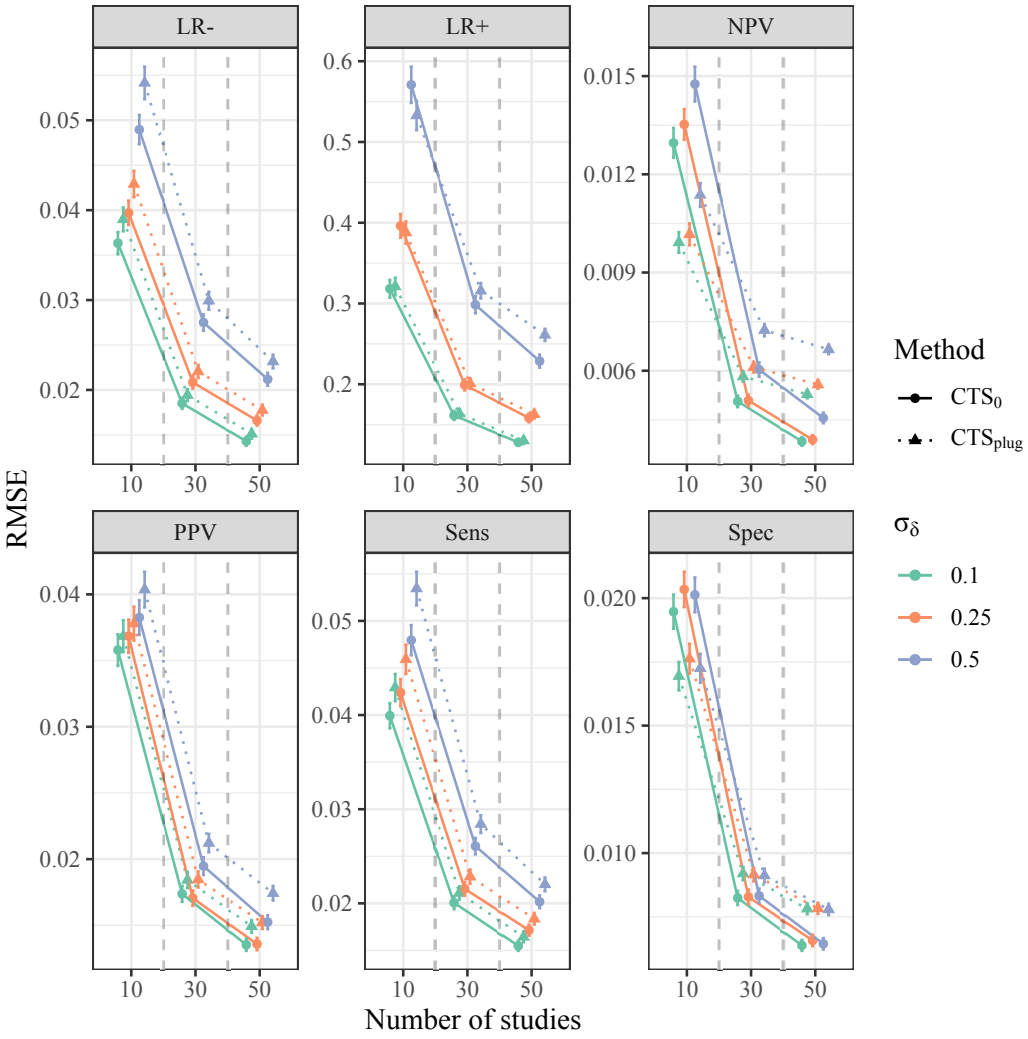

Figure 2.2: Plot of RMSE for CTS_0 (solid lines) and CTS_{plug} (dotted lines). Each panel plots RMSE for a different CTS against sample size $\in \{10, 30, 50\}$ on the x-axis and $\sigma_\delta \in \{0.1, 0.25, 0.5\}$ indicated by green, red, and blue lines, respectively. Vertical error bars plot $\pm 1.96 \times$ MCSE.

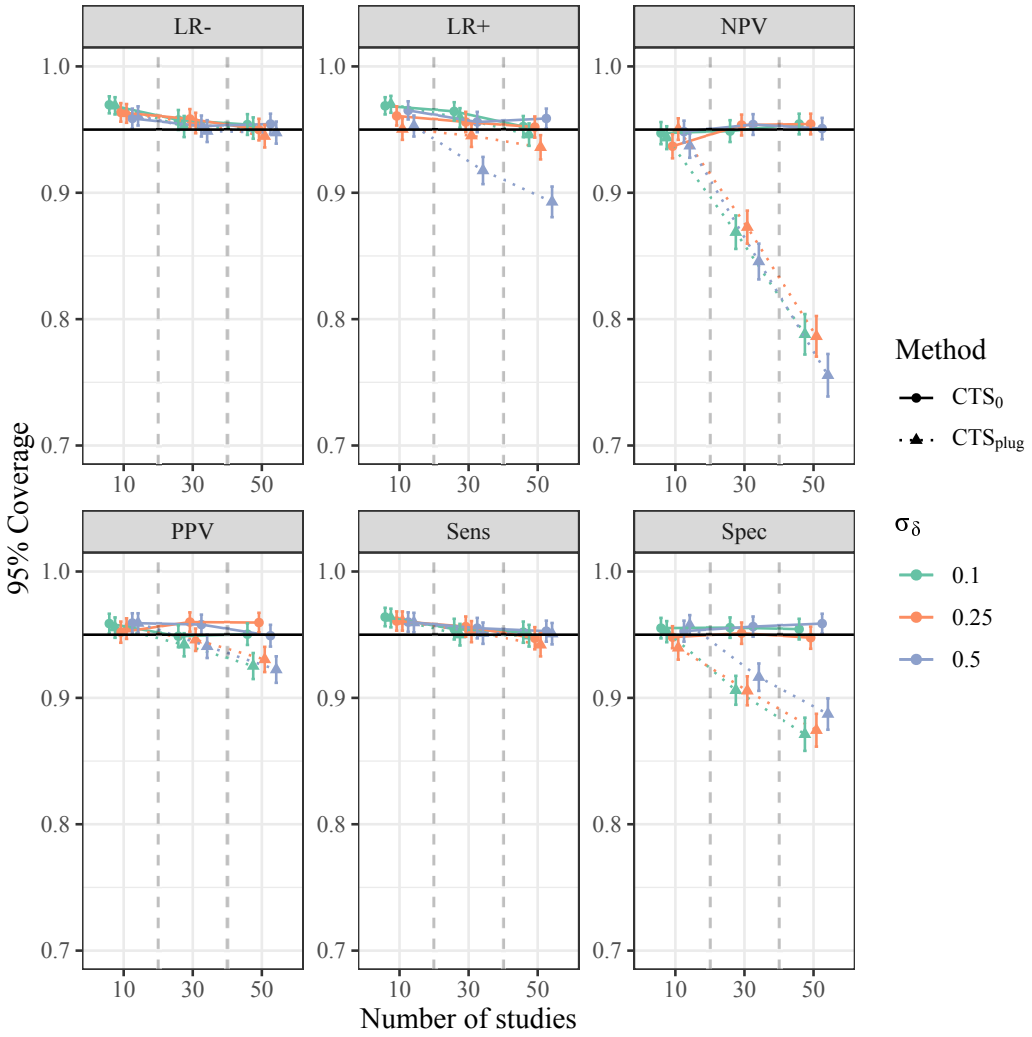

Figure 2.3: Coverage probabilities for 95% posterior intervals (PIs) for CTS_0 (solid lines) and CTS_{plug} (dotted lines) methods. Each panel plots coverage probability for different a different CTS against sample size $\in \{10, 30, 50\}$ on the x-axis, with $\sigma_\delta \in \{0.1, 0.25, 0.5\}$ indicated by green, red, and blue lines respectively. Vertical bars plot $\pm 1.96 \times$ MCSE.

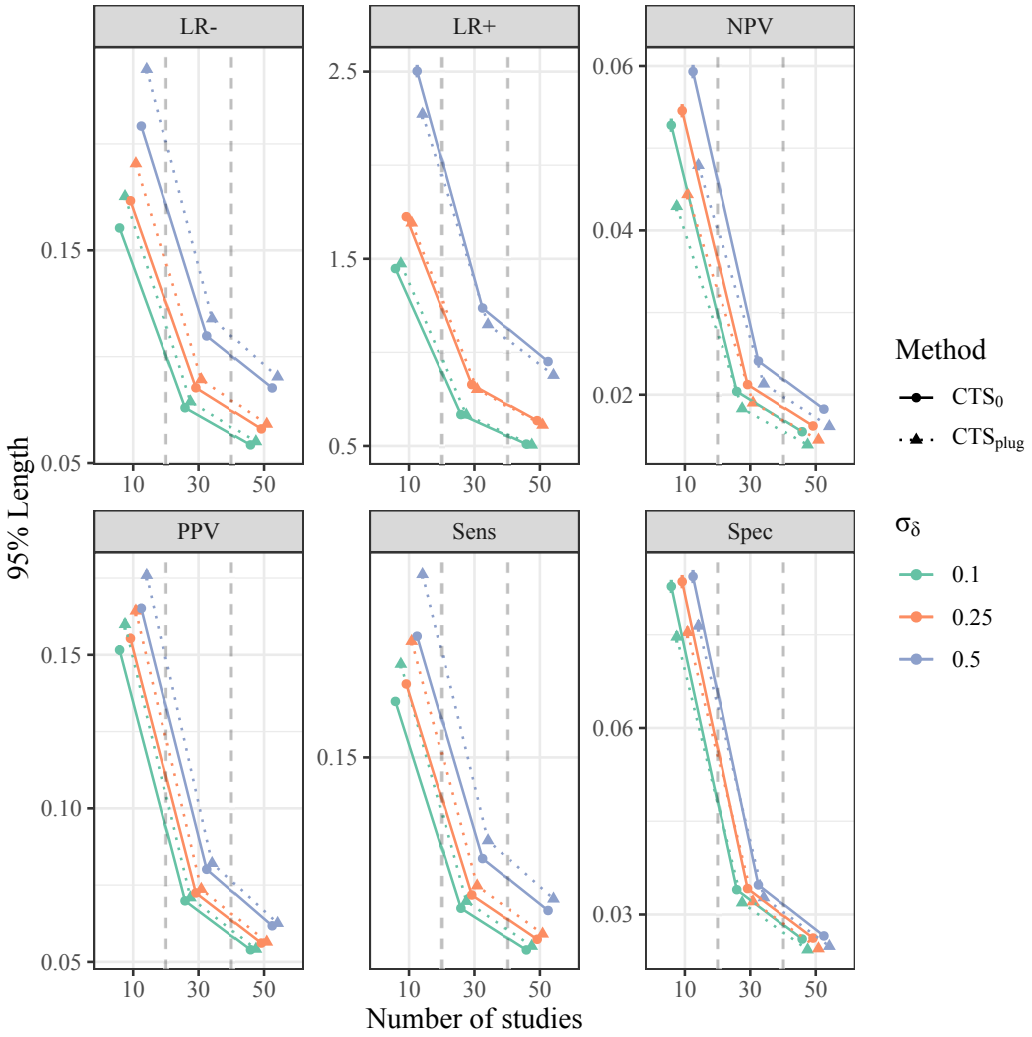

Figure 2.4: Average 95% posterior interval (PI) lengths for CTS_0 (solid lines) and CTS_{plug} (dotted lines). PI endpoints were taken as the 2.5^{th} and 97.5^{th} posterior quantiles. Each panel plots average 95% PI length for a different CTS against sample size $\in \{10, 30, 50\}$ on the x-axis, and $\sigma_\delta \in \{0.1, 0.25, 0.5\}$ indicated by green, red, and blue lines respectively. Vertical bars, though difficult to see for most points, plot $\pm 1.96 \times$ MCSE.

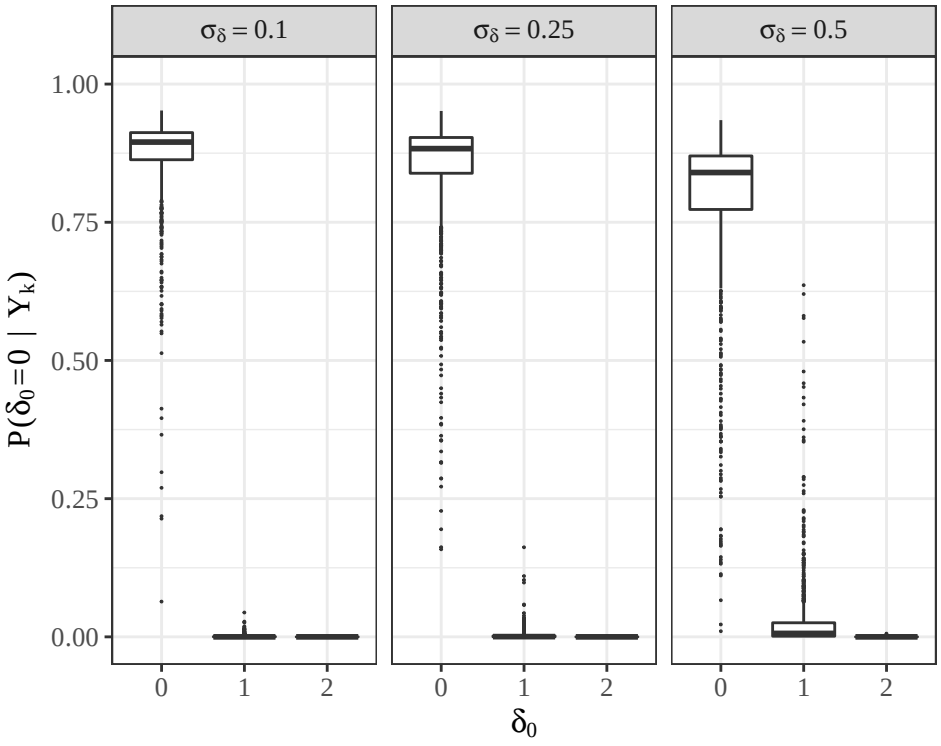

Figure 2.5: Boxplots of posterior probability $P(\delta_0 = 0|Y_k)$ that the population log odds ratio is zero for each combination of σ_δ and δ_0 in simulation 3.

Table 2.4: Results of 20 meta-analyses of syncope studies that had posterior mean $P(\delta_0 = 0|Y) < 0.5$. Each row is for a different risk factor (RF). The second column lists the number of studies that reported a 2×2 table of bad outcomes by presence/absence of the RF. Column 3 is the posterior probability $P(\delta_0 = 0|Y)$ from the 3RE-SAS model, sorted from lowest to highest. Columns 4-7 list posterior means and 95% CIs for $LR+_0$, $LR-_0$, NPV_0, and PPV_0 for each RF. Abbreviations: CHF = congestive heart failure; ECG = abnormal electrocardiogram; Resp. Rate = respiratory rate; CVD = cerebrovascular disease.

RF	Num. Papers	Spike	LR+	LR-	NPV	PPV
Male Gender	7	0.000	1.39 (1.29, 1.50)	0.72 (0.65, 0.79)	0.93 (0.83, 0.97)	0.12 (0.06, 0.24)
Age	6	0.001	2.08 (1.66, 2.65)	0.44 (0.32, 0.60)	0.94 (0.85, 0.98)	0.17 (0.08, 0.34)
CHF	8	0.004	3.42 (2.44, 4.80)	0.82 (0.74, 0.89)	0.89 (0.79, 0.95)	0.26 (0.15, 0.42)
Heart Disease	9	0.007	2.25 (1.70, 2.94)	0.79 (0.69, 0.88)	0.92 (0.84, 0.96)	0.17 (0.10, 0.29)
Arrhythmia	6	0.013	3.06 (2.06, 4.34)	0.78 (0.64, 0.90)	0.93 (0.84, 0.97)	0.18 (0.09, 0.33)
Dyspnea	6	0.023	2.78 (1.87, 4.48)	0.88 (0.79, 0.94)	0.87 (0.78, 0.92)	0.28 (0.18, 0.41)
White Race	3	0.050	1.26 (1.10, 1.56)	0.61 (0.46, 0.81)	0.94 (0.82, 0.98)	0.10 (0.04, 0.27)
Troponin	3	0.052	3.88 (1.96, 7.67)	0.63 (0.38, 0.88)	0.95 (0.90, 0.97)	0.21 (0.13, 0.32)
ECG	6	0.067	2.33 (1.54, 3.62)	0.60 (0.40, 0.85)	0.91 (0.83, 0.96)	0.24 (0.14, 0.38)
Resp. Rate	2	0.078	3.69 (1.86, 6.95)	0.93 (0.84, 0.98)	0.90 (0.69, 0.97)	0.24 (0.09, 0.51)
Creatinine	2	0.124	4.76 (2.06, 12.40)	0.86 (0.68, 0.96)	0.91 (0.70, 0.97)	0.24 (0.09, 0.50)
Urea	2	0.128	6.62 (2.61, 17.69)	0.83 (0.63, 0.96)	0.90 (0.69, 0.97)	0.29 (0.11, 0.57)
CVD	3	0.154	1.71 (1.27, 2.34)	0.92 (0.83, 0.97)	0.95 (0.82, 0.98)	0.07 (0.03, 0.23)
Trauma	2	0.165	2.07 (1.21, 3.46)	0.81 (0.56, 0.97)	0.93 (0.85, 0.96)	0.14 (0.08, 0.27)
Hematocrit	7	0.194	2.56 (1.48, 4.69)	0.92 (0.82, 0.98)	0.87 (0.79, 0.93)	0.23 (0.14, 0.36)
Diabetes	6	0.219	1.71 (1.25, 2.30)	0.88 (0.78, 0.96)	0.93 (0.83, 0.97)	0.12 (0.06, 0.25)
Hypotension	5	0.233	4.95 (1.95, 13.94)	0.91 (0.80, 0.98)	0.86 (0.73, 0.93)	0.31 (0.17, 0.49)
Hypertension	5	0.296	1.42 (1.07, 1.88)	0.72 (0.50, 0.99)	0.93 (0.78, 0.98)	0.12 (0.05, 0.30)
Murmur	4	0.299	2.22 (1.16, 5.51)	0.90 (0.72, 1.00)	0.84 (0.69, 0.92)	0.25 (0.13, 0.46)
Pacemaker	3	0.329	2.30 (1.21, 3.78)	0.95 (0.89, 0.99)	0.87 (0.65, 0.96)	0.21 (0.07, 0.46)

Table 2.5: Results of 20 meta-analyses of syncope studies that had posterior mean $P(\delta_0 = 0|Y) < 0.5$. Each row is for a different risk factor (RF). The second column lists the number of studies that reported a 2×2 table of bad outcomes by presence/absence of the RF. Column 3 is the posterior probability $P(\delta_0 = 0|Y)$ from the 3RE-SAS model, sorted from lowest to highest. Columns 4-5 list posterior means and 95% CIs for Sens_0 and Spec_0 for each RF. Abbreviations: CHF = congestive heart failure; ECG = abnormal electrocardiogram; Resp. Rate = respiratory rate; CVD = cerebrovascular disease.

RF	Num. Papers	Spike	Sens	Spec
Male Gender	7	0.000	0.58 (0.54, 0.62)	0.58 (0.56, 0.60)
Age	6	0.001	0.71 (0.58, 0.82)	0.62 (0.51, 0.72)
CHF	8	0.004	0.24 (0.16, 0.32)	0.93 (0.90, 0.95)
Heart Disease	9	0.007	0.33 (0.24, 0.44)	0.84 (0.78, 0.88)
Arrhythmia	6	0.013	0.30 (0.17, 0.47)	0.88 (0.76, 0.94)
Dyspnea	6	0.023	0.18 (0.12, 0.28)	0.93 (0.89, 0.95)
White Race	3	0.050	0.77 (0.53, 0.89)	0.36 (0.20, 0.59)
Troponin	3	0.052	0.47 (0.22, 0.74)	0.79 (0.51, 0.93)
ECG	6	0.067	0.57 (0.41, 0.73)	0.70 (0.58, 0.80)
Resp. Rate	2	0.078	0.10 (0.05, 0.20)	0.97 (0.95, 0.98)
Creatinine	2	0.124	0.18 (0.08, 0.36)	0.95 (0.91, 0.97)
Urea	2	0.128	0.21 (0.09, 0.40)	0.96 (0.92, 0.98)
CVD	3	0.154	0.17 (0.10, 0.31)	0.90 (0.79, 0.94)
Trauma	2	0.165	0.34 (0.11, 0.68)	0.78 (0.43, 0.95)
Hematocrit	7	0.194	0.14 (0.07, 0.26)	0.93 (0.87, 0.96)
Diabetes	6	0.219	0.26 (0.17, 0.37)	0.84 (0.77, 0.89)
Hypotension	5	0.233	0.12 (0.06, 0.24)	0.96 (0.92, 0.98)
Hypertension	5	0.296	0.60 (0.43, 0.75)	0.55 (0.41, 0.68)
Murmur	4	0.299	0.20 (0.07, 0.46)	0.87 (0.65, 0.96)
Pacemaker	3	0.329	0.09 (0.05, 0.15)	0.96 (0.94, 0.97)

Table 2.6: Results of 11 meta-analyses of syncope studies that had posterior mean $P(\delta_0 = 0|Y) > 0.5$. Each row is for a different risk factor (RF). The second column lists the number of studies that reported a 2×2 table of bad outcomes by presence/absence of the RF. Column 3 is the posterior probability $P(\delta_0 = 0|Y)$ from the 3RE-SAS model, sorted from lowest to highest. Columns 4-7 list posterior means and 95% CIs for $LR+_0$, $LR-_0$, NPV_0, and PPV_0 for each RF. Abbreviations: Arr. Rx = arrhythmic medication; Prev. Syncope = previous syncope.

RF	Num. Papers	Spike	LR+	LR-	NPV	PPV
Oxygen	2	0.512	1.63 (0.98, 2.57)	0.88 (0.71, 1.01)	0.89 (0.81, 0.94)	0.17 (0.10, 0.29)
Seizure	3	0.646	1.69 (0.79, 3.97)	0.97 (0.88, 1.01)	0.92 (0.74, 0.98)	0.09 (0.02, 0.30)
Arr. Rx	2	0.674	2.61 (0.59, 10.58)	0.97 (0.82, 1.05)	0.81 (0.64, 0.91)	0.24 (0.10, 0.44)
Effort	3	0.716	1.70 (0.81, 3.22)	0.95 (0.86, 1.02)	0.87 (0.77, 0.92)	0.19 (0.10, 0.32)
Prev. Syncope	3	0.749	0.93 (0.42, 2.18)	1.08 (0.93, 1.40)	0.84 (0.66, 0.92)	0.12 (0.05, 0.27)
Chest Pain	3	0.750	1.67 (0.69, 3.98)	0.97 (0.86, 1.04)	0.85 (0.69, 0.93)	0.18 (0.08, 0.35)
Supine	2	0.761	2.14 (0.72, 9.49)	1.35 (0.56, 2.33)	0.82 (0.59, 0.93)	0.23 (0.09, 0.49)
Palpitations	4	0.774	1.82 (0.75, 4.91)	0.96 (0.83, 1.03)	0.83 (0.69, 0.91)	0.22 (0.11, 0.41)
Stroke	2	0.790	1.01 (0.38, 2.61)	1.07 (0.88, 1.45)	0.89 (0.77, 0.95)	0.08 (0.03, 0.20)
No Prodromes	6	0.810	1.24 (0.86, 1.86)	0.94 (0.70, 1.26)	0.90 (0.82, 0.94)	0.12 (0.07, 0.21)
Hispanic	2	0.839	1.01 (0.51, 2.07)	1.01 (0.90, 1.09)	0.89 (0.74, 0.95)	0.10 (0.04, 0.24)

Table 2.7: Results of 11 meta-analyses of syncope studies that had posterior mean $P(\delta_0 = 0|Y) > 0.5$. Each row is for a different risk factor (RF). The second column lists the number of studies that reported a 2×2 table of bad outcomes by presence/absence of the RF. Column 3 is the posterior probability $P(\delta_0 = 0|Y)$ from the 3RE-SAS model, sorted from lowest to highest. Columns 4-5 list posterior means and 95% CIs for Sens_0 and Spec_0 for each RF. Abbreviations: Arr. Rx = arrhythmic medication; Prev. Syncope = previous syncope.

RF	Num. Papers	Spike	Sens	Spec
Oxygen	2	0.512	0.27 (0.13, 0.52)	0.82 (0.59, 0.91)
Seizure	3	0.646	0.08 (0.03, 0.21)	0.95 (0.86, 0.98)
Arr. Rx	2	0.674	0.09 (0.03, 0.26)	0.94 (0.85, 0.97)
Effort	3	0.716	0.12 (0.06, 0.23)	0.92 (0.87, 0.95)
Prev. Syncope	3	0.749	0.16 (0.04, 0.44)	0.81 (0.52, 0.94)
Chest Pain	3	0.750	0.11 (0.05, 0.22)	0.92 (0.89, 0.94)
Supine	2	0.761	0.38 (0.11, 0.71)	0.68 (0.35, 0.90)
Palpitations	4	0.774	0.12 (0.04, 0.32)	0.91 (0.75, 0.97)
Stroke	2	0.790	0.16 (0.03, 0.50)	0.82 (0.48, 0.95)
No Prodromes	6	0.810	0.45 (0.27, 0.64)	0.60 (0.42, 0.75)
Hispanic	2	0.839	0.12 (0.06, 0.26)	0.88 (0.78, 0.92)

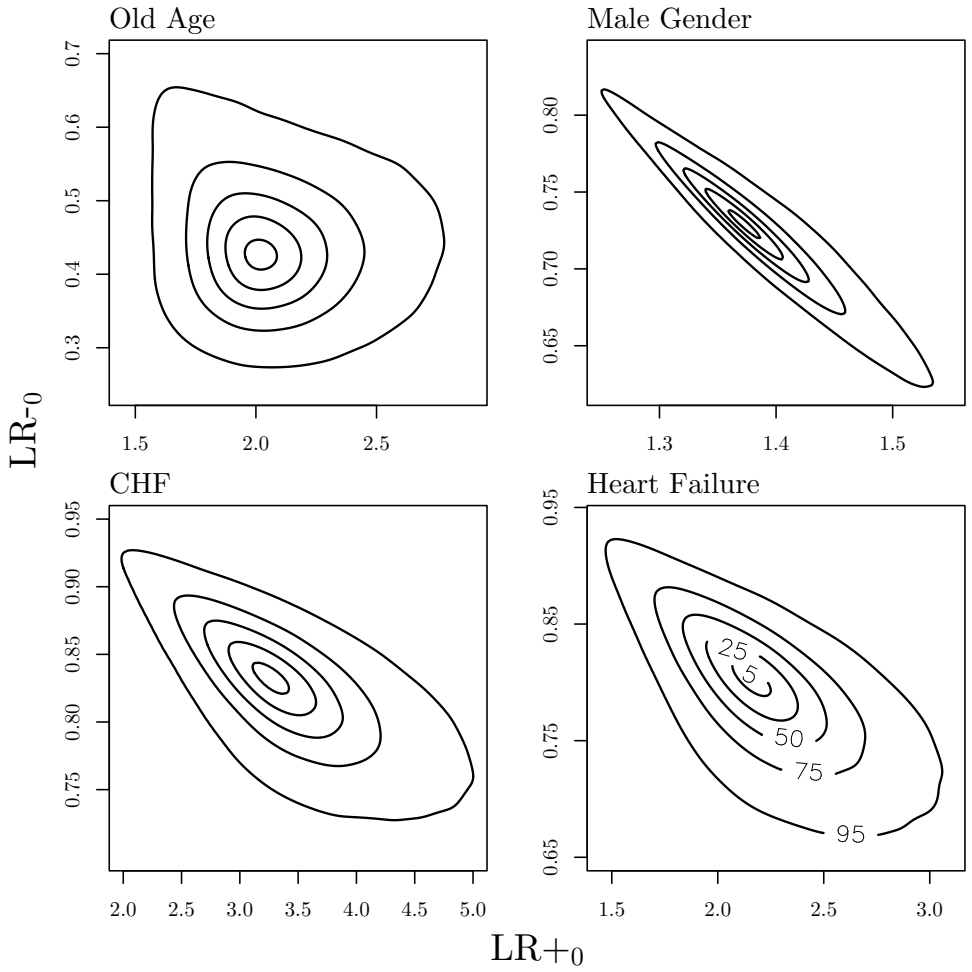

Figure 2.6: Posterior contour plots of LR− on the y-axis against LR+ on the x-axis for the four syncope risk factors with the smallest posterior $P(\delta_0 = 0|Y)$. These are old age, male gender, history of congestive heart failure (CHF), and history of heart disease. Higher (lower) values of LR+ (LR−) signal stronger diagnostic utility. Contour lines represent 5%, 25%, 50%, 75%, and 95% credible regions.

P(RF)	LR-	LR+	NPV	PPV	Sens	Spec
0.05	0.82 (0.67, 0.93)	5.63 (2.77, 11.25)	0.95 (0.89, 0.97)	0.21 (0.13, 0.33)	0.21 (0.12, 0.36)	0.96 (0.95, 0.96)
0.10	0.71 (0.53, 0.88)	4.47 (2.42, 7.78)	0.95 (0.89, 0.97)	0.21 (0.13, 0.33)	0.35 (0.21, 0.52)	0.92 (0.90, 0.93)
0.25	0.53 (0.33, 0.82)	2.86 (1.80, 4.08)	0.95 (0.89, 0.97)	0.21 (0.13, 0.33)	0.59 (0.39, 0.74)	0.78 (0.76, 0.81)

Table 2.8: Posterior summaries for the RF Troponin given known values of P(RF)$\in \{0.05, 0.10, 0.25\}$ reported as mean (95% CI).

CHAPTER 3

Mitigating Publication Bias Using Bayesian Stacking

Results from a meta-analysis may be skewed and unreliable in the presence of publication bias, where the publication or non-publication of a study depends on the statistical significance or magnitude of its results (Rothstein et al., 2006). Various statistical methods have been proposed for meta-analysis with suspected publication bias, including hypothesis tests for the presence/magnitude of publication bias, methods for calculating bias-corrected parameter estimates, and sensitivity analyses that use a grid representing varying levels of publication bias and estimate parameters of interest under each assumed scenario. If results do not change much under an assumption of severe publication bias they are robust, and if results do change under an assumption of mild publication bias they are sensitive. Most methods are either based on the funnel plot or selection models.

Methods based on the funnel plot (Light and Pillemer, 1984) – a scatterplot of effect sizes on the x-axis against their standard errors on the $y-$axis – inspect the plot's asymmetry to test or correct for bias. Say we have S studies indexed by $i = 1, \ldots, S$ with estimated effect sizes y_i and associated standard errors s_i. A popular non-parametric test for publication bias (Begg and Mazumdar, 1994) defines

standardized effect sizes y_i^* as

$$\begin{aligned} y_i^* &= (y_i - \overline{y})/(v_i^*)^{1/2} \\ \overline{y} &= \Big(\sum_j (v_i^{-1}) y_i\Big) / \Big(\sum_j v_i^{-1}\Big) \\ v_i^* &= v_i - \Big(\sum_j v_i^{-1}\Big)^{-1}, \end{aligned}$$

where $\overline{y}$ is the inverse-variance weighted mean effect size and v_i^* is the variance of $y_i - \overline{y}$. Begg and Mazumdar (1994) measure the rank correlation between pairs (y_i^*, v_i) with Kendall's tau, where a symmetric funnel plot would have correlation near zero. Egger's test (Egger et al., 1997) fits a linear regression of standardized effect sizes y_i/s_i against the inverse standard errors $1/s_i$, i.e.

$$y_i/s_i = \alpha + \beta \times (1/s_i) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2). \tag{3.1}$$

Egger et al. (1997) test the null hypothesis $H_0\colon \alpha = 0$; justification for testing $\alpha = 0$ is that, on a plot with $1/s_i$ on the x-axis and y_i/s_i on the y-axis, small studies will have small $1/s_i$ and small y_i/s_i and will thus be close to zero on both $x-$ and $y-$axes, while large studies will have large $1/s_i$ and large y_i/s_i if the treatment is effective and will be far from zero on both $x-$ and $y-$axes. So a set of studies from a homogeneous population with no publication bias will produce a regression line running through the origin at $y/s = 0$. Macaskill et al. (2001) proposed a variant of Egger's test for binary outcome data comparing a treatment with placebo, regressing log-odds ratios y_i on total sample sizes n_i and weighting observations by the inverse variance of the pooled log-odds of the event in study i. Peters et al. (2006) proposed a weighted regression similar to Macaskill et al. (2001), but instead regresses effects y_i on the

reciprocal of the sample size. Lin and Chu (2018) develop a measure for the severity of publication bias based on the skewness of standardized effects y_i/s_i.

Another funnel plot-based method, the trim-and-fill method (Duval and Tweedie, 2000), calculates a bias-adjusted estimate for the mean effect θ by 1) estimating the number of missing studies k_0, 2) "trimming" (removing) the smaller studies that are causing funnel plot asymmetry, 3) estimating the mean effect θ with the remaining studies, and 4) replacing trimmed studies, imputing their missing counterparts to recreate a symmetric funnel plot, and re-estimating the mean effect and its variance. Duval and Tweedie (2000) recommend using trim-and-fill as a sensitivity analysis based on the *potential* number of missing studies, with general guidelines for sensitivity analysis for trim-and-fill given in Shi and Lin (2019). The trim-and-fill method is the only funnel plot-based method that offers an adjusted mean estimate, and it is not recommended in a random-effects meta-analysis (Jin et al., 2015).

A second class of methods are based on *selection models*, first described in Hedges (1984). Let Y be a random variable of effect sizes for studies in a population and let Θ be the parameters determining the sampling density $f_Y(y;\Theta)$. Selection models assume a biased sampling scheme where only a subset of all studies in the population are included in a meta-analysis, and the probability of a study being observed (published) is given by a weight function $w(y;\lambda)$, where λ, a scalar or vector parameter, determines how certain studies may be more or less likely to be published. The observed effects y_i, $i = 1, \ldots, S$ come from the weighted density

$$f^*(y_i;\Theta,\lambda) = \frac{f(y_i;\Theta)w(y_i;\lambda)}{\int f(y;\Theta)w(y;\lambda)dy} \tag{3.2}$$

and the likelihood function for Θ based on the observed studies is

$$L(\Theta, \lambda) = \prod_{i=1}^{S} f^*(y_i; \Theta, \lambda). \tag{3.3}$$

Some selection models explicitly model the probability of publication for individual studies as a function of their p-values (Iyengar and Greenhouse, 1988; Hedges, 1992; Givens et al., 1997; Vevea and Hedges, 1995) or as a function of both the effect size and standard error (Copas, 1999; Copas and Shi, 2000, 2001). Hedges (1992) introduced stepped weight functions of p-values by dividing the unit interval [0, 1] into K segments with $K - 1$ change points, where studies that have p-values in different segments have different probabilities of publication. We refer to stepped selection functions by the number of change points, i.e. a 1-step selection function has a single change point at possibly $p = 0.05$, or a 2-step function might have change points at $p = 0.05, 0.10$. Selection models have been recommended primarily for sensitivity analyses because of identifiability issues in smaller meta-analyses (Vevea and Woods, 2005; Jin et al., 2015). However, Bayesian implementations of the Copas selection model (Mavridis et al., 2013; Bai et al., 2020) have been proposed for estimation of mean effect sizes.

Recent approaches to mitigating publication bias have used Bayesian model averaged meta-analysis (BMA-MA) to consider a set of potential selection functions. Guan and Vandekerckhove (2016) consider four different selection functions of p-values, including a no-bias model, an extreme-bias model where studies with p-values $p > \alpha$ are never published, a 1-step function where studies with $p > \alpha$ are published with some probability $\pi < 1$ and studies with $p < \alpha$ are published with probability 1, and a model where the probability of publication decreases exponentially with

p. Guan and Vandekerckhove (2016) only implement the models in a fixed-effects framework. Maier et al. (2022) evaluates a set of 12 models, using a $2 \times 2 \times 3$ factorial design with fixed/random effects, a true null/alternative hypothesis, and the presence/absence of publication bias with two possible selection functions. Maier et al. (2022) fit one-step and two-step selection functions based on p-values when publication bias is assumed, where the probability of publication changes at $p = 0.05$ (one-step) or at both $p = 0.05$ and $p = 0.10$ (two-step). They call their method Robust Bayesian Meta-analysis (RoBMA).

Bayesian model averaging (BMA) effectively assumes that one of the considered models is the "true" model, which is called the $\mathcal{M}$-closed setting (Bernardo and Smith, 2009). BMA does not perform as well under the $\mathcal{M}$-complete or $\mathcal{M}$-open settings, where the true data generating mechanism is too complex to implement or to put into a probabilistic framework (Bernardo and Smith, 2009; Le and Clarke, 2017). Multiple issues arise for BMA in these settings, including (a) the need to specify prior model probabilities, which makes little sense when we know the true model is not in our list, and (b) the model weights from BMA will converge to 1 for the model "closest" to the true model in terms of Kullback-Leibler divergence, and 0 for all others (Clyde and Iversen, 2013). Yao et al. (2018, 2021) proposed *Bayesian stacking of predictive distributions* as a method for model combination and showed that it outperforms BMA in a variety of $\mathcal{M}$-complete and $\mathcal{M}$-open settings and avoids issues (a) and (b) above. Given data $y = (y_1, \ldots, y_S)$ and K candidate models $M_1, \ldots, M_K$, the goal is to find a predictive distribution that is close to the unknown true data generating mechanism. Yao et al. (2018) suggest a weighted average of model-specific posterior predictive distributions and find model weights $r = (r_1, \ldots, r_K)$ in the K-simplex $r \in \mathcal{R}_1^K = \{r \in [0,1]^K : \sum_{k=1}^K r_k = 1\}$. They do

this by maximizing the expected log-predictive density (elpd) of the weighted leave-one-out (LOO) predictive densities $p(y_i|y_{-i}, M_k)$ evaluated at y_i, where y_{-i} is the data y with observation i left out. It would be computationally costly to refit each model M_k S times, so LOO densities $p(y_i|y_{-i}, M_k)$ are approximated using Pareto-smoothed importance sampling (Vehtari et al., 2017). We explain Bayesian stacking in detail in Section 3.1.1.

Given that the true data generating mechanism for publication bias is likely much too complex to be specified in a simple selection model, we propose using Bayesian stacking to mitigate publication bias by fitting multiple Bayesian selection models and stacking over them. Models of publication bias that poorly predict the observed data with LOO cross validation will be given little weight. We propose stacking over multiple types of models, including step functions (Vevea and Hedges, 1995) Bayesian Copas selection models (Mavridis et al., 2013; Bai et al., 2020), and a novel sloped selection model based on p-values. Section 3.1 describes Bayesian stacking, the selection models for publication bias that we use, and how to implement Bayesian stacking of selection models. We then describe and summarize a simulation study in Section 3.2. The purpose of the simulation is to compare a stacked estimate of the mean effect size to estimates from individual selection models and RoBMA when the true data generator is not one of the fitted selection models. We use the stacked model to adjust for publication bias for three datasets on (1) the effects of second-hand smoke on the likelihood of developing lung cancer, (2) gender effects in grant proposals, and (3) the effects of cognitive behavioral therapy on recidivism in Section 3.3. The paper closes with discussion.

3.1 Methods

We are doing a meta-analysis with S studies indexed by $i = 1, \ldots, S$, where each study provides an estimated effect y_i and standard error s_i. We calculate study i's 1-sided p-value as $p(y_i, s_i) = 1 - \Phi(y_i/s_i)$ and study i's 2-sided p-value as $p(y_i, s_i) = 2 \times \big(1 - \Phi(|y_i|/s_i)\big)$ where $\Phi(\cdot)$ is the standard normal cumulative distribution function. The sampling density for study i is

$$y_i = \theta_i + s_i \epsilon_i \tag{3.4}$$

$$\theta_i | \theta, \tau^2 \sim \mathrm{N}(\theta, \tau^2) \tag{3.5}$$

where θ_i is a random study effect normally distributed around global mean θ with unknown variance τ^2, and $\epsilon_i \sim \mathrm{N}(0, 1)$ is a random residual. We marginalize over random effects θ_i in model (3.4) - (3.5) giving a marginal model for y_i as

$$y_i | \theta, \tau^2 \sim \mathrm{N}(\theta, s_i^2 + \tau^2). \tag{3.6}$$

We will fit K separate selection models M_k, $k = 1, \ldots, K$ to the S studies in the analysis. No model is likely to capture the true data generating mechanism, so we will stack over the K models to find a predictive distribution closer to the true data generating mechanism. We describe Bayesian stacking before defining the models that we stack over.

3.1.1 Bayesian stacking

We have K candidate models M_k indexed by $k = 1, \ldots, K$ and data $y = \{y_i\}$, where each model M_k includes a common parameter θ. We want to build a more robust model to make inference about θ by combining the model-specific posteriors $P_k(\theta|y)$, where $P_k(\theta|y)$ is the posterior distribution of θ under model M_k. The $\mathcal{M}$-open perspective assumes our list of candidate models does not include the true data generating mechanism (Bernardo and Smith, 2009). Bayesian stacking (Yao et al., 2018) is an alternative to Bayesian model averaging that has been shown to have superior performance in several $\mathcal{M}$-open data analysis scenarios.

Bayesian stacking utilizes *proper scoring rules* (Gneiting and Raftery, 2007), which measure the concordance of a probabilistic forecast P over a sample space Y with the true data-generating mechanism Q over Y. Formally, say Y is a sample space on $[-\infty, \infty]$; for a probabilistic forecast P over Y, identified by its density function $p(y), y \in Y$, the log-score S is defined as $S(P, y) = \log(p(y))$ for a realization y from Y. The expected score of a forecast P under the *true* sampling density Q is

$$S(P, Q) = \int S(P, y) dQ(y). \tag{3.7}$$

The generic stacking problem is to find the optimal vector of model weights $r = (r_1, \ldots, r_K)$ in the K-simplex that maximizes the expected log-score of the weighted sum of predictive distributions $\sum_{k=1}^{K} r_k p(\tilde{y}|y, M_k)$ of future data $\tilde{y}$ relative to the *hypothetical* true distribution of future data $p_T(\tilde{y}|y)$. Yao et al. (2018) define the

stacking problem as solving

$$\arg\max_{r\in\mathcal{R}_1^K} S\Big(\sum_{k=1}^{K} r_k p(\tilde{y}|y, M_k), p_T(\tilde{y}|y)\Big) \tag{3.8}$$

for r. Because we do not know the true distribution $p_T(\tilde{y}|y)$, Yao et al. (2018) replace the "true" predictive distribution $p_T(\tilde{y}|y)$ with the empirical CDF $\hat{F}_n(x) = \frac{1}{n}\sum_i \mathbb{1}_{[y_i<x]}$, and replace model k's predictive distribution $p(\tilde{y}|y, M_k)$ with its corresponding leave-one-out (LOO) predictive distribution

$$\hat{p}_{k,-i}(y_i) = \int p(y_i|\theta_k, M_k)p(\theta_k|y_{-i}, M_k)d\theta_k, \tag{3.9}$$

where θ_k are the parameters in model k and subscript $-i$ denotes the data y with observation i left out. The expected log-score is then a sum over the n data points y_i rather than an integral over the true sampling distribution. The stacking problem with $S(P, y) = \log(y)$ reduces to solving for weights r with

$$(\hat{r}_1, \ldots, \hat{r}_K) = \arg\max_{r\in\mathcal{R}_1^K} \frac{1}{n}\sum_{i=1}^{n} S\Big(\sum_{k=1}^{K} r_k \hat{p}_{k,-i}, y_i\Big) \tag{3.10}$$

$$= \arg\max_{r\in\mathcal{R}_1^K} \frac{1}{n}\sum_{i=1}^{n} \log\sum_{k=1}^{K} r_k \hat{p}_{k,-i}(y_i). \tag{3.11}$$

The *stacked posterior distribution* of a common parameter θ is the mixture $P(\theta|y, \hat{r}) = \sum_k \hat{r}_k P_k(\theta|y)$, and samples from the stacked posterior are obtained by mixing $\hat{r}_k \times T$ random samples from each model k's posterior distribution $P_k(\theta|y)$ and combining them for a total of T posterior samples.

Rather than refitting each model M_k S times, Bayesian stacking uses Pareto

smoothed importance sampling (PSIS) (Vehtari et al., 2017) to calculate LOO predictive densities $\hat{p}_{k,-i}(y_i)$. The PSIS calculation provides a diagnostic value $\hat{h}_{ki}$ for each LOO approximation $\hat{p}_{k,-i}(y_i)$ which measures the reliability of the approximation to $p_{k,-i}(y_i)$, where $\hat{h}_{ki} > 0.7$ signals a potentially unreliable approximation. For calculations of $\hat{p}_{k,-i}(y_i)$ with diagnostic $\hat{h}_{ki} > 0.7$, we refit the model to sample from the exact LOO distribution $p_{k,-i}(y_i)$. To calculate LOO predictive densities for models M_k using PSIS, we need the posterior distribution of the point-wise log-likelihood $p_k(y_i|\theta_k, M_k)$ for each observation y_i.

3.1.2 Stepped selection function of p-values

To model the process through which studies are chosen to be in a meta-analysis, define a stepped weight function $w(\cdot)$ (Vevea and Hedges, 1995; Vevea and Woods, 2005) that is constant on intervals, where $w(p)$ is the probability that a study with p-value p is observed. We divide the unit interval into J sub-intervals, where a study with a p-value p_i has a probability of publication that corresponds to the sub-interval p_i falls into. Let $a_{j-1} > a_j$, $j = 1, \ldots, J$, be decreasing change points where $a_0 = 1$ and $a_J = 0$ and define

$$w(p) = \begin{cases} \omega_1 & \text{if } a_1 < p < 1 \\ \omega_j & \text{if } a_j < p < a_{j-1} \\ \omega_J & \text{if } 0 < p < a_{J-1}. \end{cases} \tag{3.12}$$

Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_J)$ be the vector of weights associated with the J sub-intervals of $[0, 1]$. The likelihood contribution for each study i given θ, τ^2, and weight function

$w(\cdot)$ is

$$f(y_i|\theta, \tau^2, \boldsymbol{\omega}) = \frac{\phi(y_i; \theta, \tau^2 + s_i^2) \times w(p_i)}{\int \phi(x; \theta, \tau^2 + s_i^2) \times w(p(x, s_i))dx}, \tag{3.13}$$

where $\phi(x; a, b)$ is a normal probability density function evaluated at x with mean a and variance b. Maier et al. (2022) place a *cumulative-Dirichlet* prior distribution on the weights $\boldsymbol{\omega}$, by first placing a symmetric Dirichlet prior on an auxiliary $J \times 1$ vector parameter $\widetilde{\boldsymbol{\omega}}$ in the $J-$simplex,

$$\widetilde{\boldsymbol{\omega}} \sim \text{Dirichlet}(\mathbf{1}_J) \tag{3.14}$$

and setting elements ω_j of $\boldsymbol{\omega}$ to be the cumulative sum

$$\omega_j = \sum_{k=1}^{j} \widetilde{\boldsymbol{\omega}}_k, \quad j = 1, \ldots, J, \tag{3.15}$$

where $\mathbf{1}_J$ is a J-vector of 1's. This restricts $\boldsymbol{\omega}$ to have increasing probability of publication with decreasing p-values, and $\omega_J = 1$. The symmetric Dirichlet prior on $\widetilde{\boldsymbol{\omega}}$ leads to prior means $(\frac{1}{J}, \frac{2}{J}, \ldots, 1)$ for $\boldsymbol{\omega}$. Restricting $\omega_J = 1$ means each ω_j is the relative probability of publication for a study in interval j compared to a study in interval J. We consider a variety of weight functions $w(p)$ by varying both the number of intervals J and the choice of change points a_j and consider both one-sided and two-sided p-values.

3.1.3 Sloped selection function of p-values

We define a new *sloped publishing probability* $w(p)$ as a continuous non-increasing piecewise linear function of p-values that is constant on the first and last intervals.

We again divide the unit interval into J segments with $J-1$ decreasing knots a_j, $j = 1, \ldots, J-1$, with $a_0 = 1$ and $a_J = 0$, and define the publishing probability as

$$w(p_i) = \begin{cases} \omega_1 & \text{if } a_1 \leq p_i \leq 1 \\ \omega_j + \frac{\omega_{j-1}-\omega_j}{a_{j-1}-a_j}(p_i - a_j) & \text{if } a_j \leq p_i < a_{j-1} \\ \omega_{J-1} & \text{if } 0 \leq p_i < a_{J-1}. \end{cases} \tag{3.16}$$

We set $\omega_{J-1} = 1$ so that $w(p_i)$ is the probability of observing a study with p-value p_i relative to a study with p-value in the interval $[0, a_{J-1}]$. At the knots, (3.16) ensures that $w(a_j) = \omega_j$, and $w(p_i)$ decreases linearly from ω_j to ω_{j-1} when $a_j < p_i < a_{j-1}$. For study i with effect y_i, standard error s_i, and p-value p_i, the sampling density of y_i is

$$f(y_i|\theta, \tau^2, \boldsymbol{\omega}) = \frac{\phi(y_i; \theta, \tau^2 + s_i^2) \times w(p_i)}{\int \phi(x; \theta, \tau^2 + s_i^2) \times w(p(x, s_i))dx}. \tag{3.17}$$

Computation of the integral in the denominator of (3.17) is tricky, but calculation is made easier by breaking the integral into a sum of smaller integrals. For one-sided p-values, we divide the real line $[-\infty, \infty]$ into J sub-intervals defined by cut points $a_j^* = s_i \Phi^{-1}(1 - a_j)$, where $a_0^* = -\infty$ and $a_J^* = \infty$. The denominator of (3.17) can be rewritten as

$$\sum_{j=0}^{J-1} \int_{a_j^*}^{a_{j+1}^*} \phi(x; \theta, \tau^2 + s_i^2) \times w(p(x, s_i))dx \tag{3.18}$$

where each range (a_j^*, a_{j+1}^*) corresponds to the range of x values that produce p-values in the range (a_{j+1}, a_j) given standard error s_i. The first and last integrals in

the sum (3.18) are

$$\int_{-\infty}^{a_1^*} \phi(x;\theta,\tau^2+s_i^2)dx = \Phi\Big((a_1^*-\theta)/\sqrt{\tau^2+s_i^2}\Big)$$

and

$$\omega_1 \int_{a_{J-1}^*}^{\infty} \phi(x;\theta,\tau^2+s_i^2)dx = \omega_1\Big(1-\Phi\Big((a_{J-1}^*-\theta)/\sqrt{\tau^2+s_i^2}\Big)\Big)$$

and are each calculated with one evaluation of the normal CDF. The integrals in (3.18) for $j = 1,\ldots,J-2$ can be calculated using quadrature methods, but to increase computation speeds we instead use properties of the normal distribution as

$$\begin{aligned}
&\int_{a_j^*}^{a_{j+1}^*} \phi(x;\theta,\tau^2+s_i^2)w(p(x,s_i))dx \\
= &\int_{a_j^*}^{a_{j+1}^*} \phi(x;\theta,\tau^2+s_i^2)\big(\omega_{j+1}+\frac{\omega_j-\omega_{j+1}}{a_j-a_{j+1}}(1-\Phi(x/s_i)-a_{j+1})\big)dx \\
= &\big(\omega_{j+1}+\frac{\omega_j-\omega_{j+1}}{a_j-a_{j+1}}(1-a_{j+1})\big)\int_{a_j}^{a_{j+1}^*} \phi(x;\theta,\tau^2+s_i^2)dx \\
&-\Big(\frac{\omega_j-\omega_{j+1}}{a_j-a_{j+1}}\Big)\int_{a_j^*}^{a_{j+1}^*} \phi(x;\theta,\tau^2+s_i^2)\Phi(x/s_i)dx.
\end{aligned} \tag{3.19}$$

The first term in equation (3.19) is the difference of two normal CDFs, while the second integral can be rewritten with a change of variables

$$\begin{aligned}
&\int_{a_j^*}^{a_{j+1}^*} \phi(x;\theta,\tau^2+s_i^2)\Phi(x/s_i)dx \\
= &\int_{\widetilde{a}_j^*}^{\widetilde{a}_{j+1}^*} \phi(x;0,1)\Phi\Big(\frac{x\sqrt{\tau^2+s_i^2}+\theta}{s_i}\Big)dx
\end{aligned} \tag{3.20}$$

where $\widetilde{a}_j^* = (a_j^* - \theta)/\sqrt{\tau^2 + s_i^2}$ is standardized. Define $\text{BvN}(z_1, z_2; \zeta)$ as the CDF of a bivariate normal distribution with both means 0, both variances 1, and correlation ζ, evaluated at z_1 and z_2. Owen (1980) showed that the integral

$$\int_{-\infty}^{U} \phi(x; 0, 1)\Phi(c_1 + c_2 x)dx = \text{BvN}\Big(\frac{c_1}{\sqrt{1 + c_2^2}}, U; \zeta = \frac{-c_2}{\sqrt{1 + c_2^2}}\Big).$$

Thus equation (3.20) simplifies to

$$\int_{-\infty}^{\widetilde{a}_{j+1}^*} \phi(x; 0, 1)\Phi\Big(\frac{x\sqrt{\tau^2 + s_i^2} + \theta}{s_i}\Big)dx - \int_{-\infty}^{\widetilde{a}_j^*} \phi(x; 0, 1)\Phi\Big(\frac{x\sqrt{\tau^2 + s_i^2} + \theta}{s_i}\Big)dx \tag{3.21}$$

$$= \quad \text{BvN}\Big(A,\, \widetilde{a}_{j+1}^*;\, \zeta\Big) - \text{BvN}\Big(A,\, \widetilde{a}_j^*;\, \zeta\Big), \tag{3.22}$$

where

$$A = \frac{\theta/s_i}{\sqrt{1 + \frac{\tau^2 + s_i^2}{s_i^2}}}$$
$$\zeta = \frac{-(\tau^2 + s_i^2)/s_i^2}{\sqrt{1 + \frac{\tau^2 + s_i^2}{s_i^2}}}.$$

We found that using (3.21) - (3.22) to calculate the sum of integrals (3.18) results in posterior sampling speeds 3-4 times faster than calculating the integrals using quadrature methods. For two-sided p-values the integral in the denominator of (3.17) can be analogously broken into $2J - 1$ x-value ranges that produce p-values in the ranges (a_{j+1}, a_j), $j = 1, \ldots, J-1$, where $a_j^* = \Phi^{-1}(a_{J-j}/2)$ and $a_{2J-j-1}^* = s_i\Phi^{-1}(1 - a_{J-j}/2)$ for $j = 1, \ldots, J-1$, $a_0 = -\infty$, and $a_{2J-1}^* = \infty$.

We generally use the sloped weight function with two change points a_1 and a_2 so

that there is a single parameter ω in (3.16) which is the probability of publication for the largest p-value interval. We model ω as Uniform$(0, 1)$.

3.1.4 Copas selection model

The Copas selection model (Copas, 1999; Copas and Shi, 2000, 2001) models the selection probability of publication as a function of study effect y_i and inverse standard error $1/s_i$. The probability of publication is modeled with a probit model. Introduce latent variable z_i modeled as

$$z_i = \gamma_0 + \frac{\gamma_1}{s_i} + \delta_i, \tag{3.23}$$

where z_i models the publication process such that study i is selected (published) only if $z_i > 0$, $\Phi(\gamma_0)$ is the baseline probability of publication as $1/s_i$ approaches zero , γ_1 is the coefficient of $1/s_i$, and δ_i is a random normal residual. Residuals ϵ_i from (3.4) and δ_i are modeled as bivariate normal

$$\begin{pmatrix} \epsilon_i \\ \delta_i \end{pmatrix} \mid \rho \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \tag{3.24}$$

where $\text{corr}(\epsilon_i, \delta_i) = \rho$. If $\rho = 0$ then the selection process does not depend on observed effect sizes, regardless of standard errors s_i, and the estimate of θ from model (3.6) is unbiased without selection modeling. Positive values of ρ indicate that larger values of y_i, relative to their true mean θ_i, are being selected for, while negative values of ρ would show selection favoring larger negative values of y_i.

The unconditional probability that a study with standard error s_i is published is

$$P(z_i > 0 | s_i, \gamma_0, \gamma_1) = \Phi(\gamma_0 + \frac{\gamma_1}{s_i}).$$

Thus, if γ_0 is large and positive then all studies are published with high probability regardless of the value of s_i or γ_1. We restrict γ_1 to be positive under the assumption that larger studies are more likely to be published for several reasons, such as more funding, and quality of writing. Larger values of γ_1 lead to larger differences in publication probabilities for studies with different standard errors.

We consider two Bayesian adaptations of the Copas model proposed by Mavridis et al. (2013) and Bai et al. (2020). Both authors recommend a vague normal prior for θ, such as $\text{N}(0, 100)$. Bai et al. (2020) place priors directly on γ_0 and γ_1 as

$$\begin{aligned} \gamma_0 &\sim \text{Uniform}(-2, 2) \\ \gamma_1 &\sim \text{Uniform}(0, s_{\max}), \end{aligned} \tag{3.25}$$

where $s_{\max}$ is the largest observed standard error. This range of values for γ_0 and γ_1 leads to unconditional selection probabilities between $\Phi(-2) = 2.5\%$ and $\Phi(3) = 99.7\%$ by restricting most of the mass for latent variables z_i to the range $(-2, 3)$. Mavridis et al. (2013) instead places priors on the lower and upper bounds for the probability of publication, P_{low} and P_{high}, as

$$\begin{aligned} P_{\text{low}} &\sim \text{Uniform}(L_1, L_2) \\ P_{\text{high}} &\sim \text{Uniform}(U_1, U_2), \end{aligned} \tag{3.26}$$

where (L_1, L_2) and (U_1, U_2) represent plausible ranges for the probability of publica-

tion for the studies with the largest and smallest standard errors, respectively. They then transform $(P_{\text{low}}, P_{\text{high}})$ to (γ_0, γ_1) with a 1-to-1 transformation

$$\begin{aligned} \gamma_0 + \frac{\gamma_1}{s_{\max}} &= \Phi^{-1}(P_{\text{low}}) \\ \gamma_0 + \frac{\gamma_1}{s_{\min}} &= \Phi^{-1}(P_{\text{high}}) \end{aligned} \tag{3.27}$$

where $s_{\min}$ is the smallest observed standard error in the sample of S studies.

Priors (3.25) are meant to be default prior distributions, while (3.26) may require more problem-specific tuning, and the two priors lead to surprisingly different posterior distributions for the mean parameter θ.

Mavridis et al. (2013) and Bai et al. (2020) both give ρ a noninformative Uniform$(-1, 1)$ prior distribution. We also consider two more informative prior choices that may be preferred. The first is a boundary-avoiding prior for ρ, whose density goes to zero at $\rho = -1$ and $\rho = 1$,

$$\begin{aligned} \widetilde{\rho} &\sim \text{Beta}(2, 2) \\ \rho &= 2\widetilde{\rho} - 1. \end{aligned} \tag{3.28}$$

The boundary-avoiding prior is more conservative than the uniform prior and shrinks ρ towards zero, and is a preferred option to avoid over-correcting for publication bias.

The second informative prior may be used if there is strong prior evidence that publication bias is present. In a review of over 1000 meta-analyses from the Cochrane Database of Systematic Reviews, Kicinski et al. (2015) found that "outcomes favoring treatment had on average a 27% higher probability to be included than other outcomes," and also noted that meta-analyses including older studies were more

likely to have publication bias. We define a positive-leaning prior

$$\rho \sim \mathrm{N}(a_\rho, b_\rho)\mathbb{1}_{\rho\in(-1,1)} \tag{3.29}$$

where $a_\rho > 0$ and $b_\rho > 0$ are known. Prior (3.29) gives more prior mass to positive values of ρ than negative values, indicating a prior belief that publication bias favoring positive results is more likely than publication bias favoring negative results. For example, setting $a_\rho = 0.44$ and $b_\rho = 1$ gives ≈ 1.5 times the prior mass to values of $\rho \in (0, 1]$ compared to $\rho \in [-1, 0]$.

3.1.5 Stacking selection models for publication bias

It would be naive to think that any of the stepped selection functions in Section 3.1.2, sloped selection functions in Section 3.1.3, or the Copas models in Section 3.1.4 represent the true data generating mechanism for publication bias. As Bayesian stacking is designed to perform well in the event that our model list does not contain the true model, we propose stacking over both Copas models (Mavridis et al., 2013; Bai et al., 2020) and a variety of stepped and sloped selection functions to obtain a more robust posterior distribution for the mean parameter θ.

To fit the Copas models rewrite model (3.4) - (3.5), (3.23) - (3.24) as

$$\begin{pmatrix} y_i \\ z_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \theta \\ u_i \end{pmatrix}, \begin{pmatrix} \tau^2 + s_i^2 & \rho s_i \\ \rho s_i & 1 \end{pmatrix} \right) \mathbb{1}_{z_i>0}. \tag{3.30}$$

We need to calculate the log-likelihood for each observation i to stack models, and

model construction (3.30) leads to a simple form for the log-likelihood

$$\begin{aligned} L(\theta, \tau^2, \rho, \gamma_0, \gamma_1 | \{y_i, s_i, z_i\}) &= \sum_{i=1}^{S} \log[p(y_i | z_i > 0, s_i)] \\ &= \sum_{i=1}^{S} \log \left[\frac{p(z_i > 0 | y_i, s_i) f(y_i | s_i)}{p(z_i > 0 | s_i)} \right] \\ &= \sum_{i=1}^{S} \Big[\log \Phi(v_i) + \log \left(\phi(y_i; \theta, \tau^2, s_i) \right) \\ &\qquad - \log \Phi(u_i) \Big] \end{aligned} \tag{3.31}$$

where $u_i = \gamma_0 + \frac{\gamma_1}{s_i}$ is the marginal mean $\mathrm{E}[z_i | s_i]$,

$$v_i = \frac{u_i + \tilde{\rho}_i \frac{y_i - \theta}{\sqrt{\tau^2 + s_i^2}}}{\sqrt{(1 - \tilde{\rho}_i^2)}}$$

is the mean of z_i conditional on y_i and s_i divided by its conditional standard deviation $\mathrm{E}[z_i | y_i, s_i] / \sqrt{\mathrm{Var}[z_i | y_i, s_i]}$, and

$$\tilde{\rho}_i = \frac{s_i}{(\tau^2 + s_i^2)^{1/2}} \rho$$

is the correlation $\mathrm{cor}(y_i, z_i)$.

While Bai et al. (2020) use default prior distributions (3.25) for the parameters γ_0 and γ_1, Mavridis et al. (2013) instead advise meta-analysts to use expert elicitation or historical data to specify (L_1, L_2) and (U_1, U_2) in (3.26) as plausible bounds for the lower and upper probabilities of publication. To avoid the need for strictly informative prior values, we specify $(L_1, L_2) = (0, 0.5)$ and $(U_1, U_2) = (0.5, 1)$, meaning we believe the lower bound for publication probability is between 0-0.5,

and the upper bound is between 0.5-1. Mavridis et al. (2013) gives τ a half-normal prior $\tau \sim \mathrm{N}(0, 10^2)\mathbb{1}_{\tau>0}$, while Bai et al. (2020) uses a half-Cauchy prior $\tau \sim \mathrm{Cauchy}(0, 1)\mathbb{1}_{\tau>0}$. We use the half-normal and half-Cauchy prior in the Mavridis Copas and Bai Copas models, respectively. We fit the two Copas models in JAGS (Plummer et al., 2003) and include them in every analysis.

In the set of stacked models we include eight stepped models, with either two-sided or one-sided selection (Two-side or One-side) and either one or two change points,

1. **Two-side (1)**: single change point at $p = 0.05$,
2. **Two-side (2)**: two change points at $p = 0.05$ and 0.50,
3. **Two-side (3)**: two change points at $p = 0.05$ and 0.20,
4. **Two-side (4)**: two change points at $p = 0.01$ and 0.10,
5. **One-side (1)**: single change point at $p = 0.025$,
6. **One-side (2)**: two change points at $p = 0.025$ and 0.50,
7. **One-side (3)**: two change points at $p = 0.025$ and 0.10,
8. **One-side (4)**: two change points at $p = 0.005$ and 0.05.

We include six sloped selection models with either two-sided or one-sided selection (Slope-T or Slope-O) and two knots,

1. **Slope-T (1)**: knots at $p = 0.05$ and 0.95,
2. **Slope-T (2)**: knots at $p = 0.05$ and 0.50,

3. **Slope-T (3)**: knots at $p = 0.01$ and 0.10,

4. **Slope-O (1)**: knots at $p = 0.025$ and 0.50,

5. **Slope-O (2)**: knots at $p = 0.025$ and 0.10,

6. **Slope-O (3)**: knots at $p = 0.005$ and 0.05.

We fit the stepped and sloped selection models in Stan (Gelman et al., 2015) because of the ability to code the custom probability distribution (3.13), which also makes calculation of the log-likelihood simple. We include the 8 stepped selection models in all simulations and data analyses. Sloped selection models take much longer to fit than stepped models, so we omit them in the main simulation in Section 3.2 but include them in a smaller simulation in Section 3.2.3 and in the data analyses in Section 3.3. The two Copas models are included in every simulation and data analysis.

3.2 Simulations

The aim of the simulation is to assess how well the stacked model described in Section 3.1.5 stacks up against the individual selection models and against Robust Bayesian Meta-analysis (RoBMA) (Maier et al., 2022) in estimating the true mean effect θ. We simulate data for meta-analyses using a $2 \times 2 \times 2 \times 4$ factorial design with

- Two selection functions where one is one-sided and one is two-sided;
- There is a moderate and an extreme version for each selection function;
- Mean effect size $\theta_0 = 0.1$ or 0.5;

- A small (10), medium (20), large (40), or very large (80) number of studies S_j per meta-analysis on average;

for $j = 1, \ldots, 32$ simulation scenarios. For the each simulation scenario j we generate $\widetilde{S}_j$ studies that are filtered through the selection function for scenario j such that an average of S_j studies per analysis survive the selection mechanism and enter the analysis. We set between-study SD $\tau = 0.2$ and let study-specific standard errors s_{ij}, $i = 1, \ldots, \widetilde{S}_j$, be distributed as Uniform(0.1, 0.8). Study-specific true mean effects θ_{ij} are distributed as $\theta_{ij}|\theta_{0j}, \tau^2 \sim \mathrm{N}(\theta_{0j}, \tau^2)$, $i = 1, \ldots, \widetilde{S}_j$. We then sample $y_{ij}|\theta_{ij} \sim \mathrm{N}(\theta_{ij}, s_i^2)$ as observed study effects. Each study i has a selection probability α_{ij} as a function of its p-value $1 - \Phi(y_{ij}/s_{ij})$ and inclusion or exclusion is determined by independent Bernoulli random variables $B_{ij}|\alpha_{ij} \sim \text{Bernoulli}(\alpha_{ij})$. The selection functions are chosen deliberately such that none of the stepped selection models, sloped selection models, or Copas models can individually capture the true selection mechanism.

We stack over the two Copas models, 8 stepped selection models, the standard random effects model for $K = 11$ candidate models. With L simulation replicates, for a given model with posterior mean estimates $\hat{\theta}^{(l)}$, $l = 1, \ldots, L$, we evaluate model performance using bias calculated as bias $= \frac{1}{L}\sum_k(\hat{\theta}^{(l)} - \theta)$, 95% credible interval (CI) coverage where 95% CI endpoints are the 2.5^{th} and 97.5^{th} posterior quantiles, and root-mean squared error (RMSE) calculated as RMSE $= \sqrt{\frac{1}{L}\sum_l(\hat{\theta}^{(l)} - \theta)^2}$.

3.2.1 Selection functions

Figure 3.1 shows moderate (M) and extreme (E) forms for the two selection mechanisms (SMs). The two SMs are deliberately designed such that none of the selection

models fitted to a selected dataset can capture the true SM. SM1 is a function of a one-sided p-value, has non-increasing selection probabilities with increasing p, and has the general form

$$f_1(p) = \begin{cases} 1 & \text{if } 0 \leq p < 0.005 \\ \exp(c_1 \times p) & \text{if } 0.005 \leq p < 0.2 \\ \exp(c_2 \times p) & \text{if } 0.2 \leq p < 0.5 \\ c_3 & \text{if } 0.5 \leq p \leq 1 \end{cases} \tag{3.32}$$

where $c = (c_1, c_2, c_3) = (-0.5, -1, 0.5)$ for f_{1M} and $c = (-2, -4, 0.1)$ for f_{1E}. SM1 is constant with selection probability 1 on the interval $[0, 0.005)$, exponential decay with different rates on $[0.005, 0.2)$ and $[0.2, 0.5)$, and constant selection probability on $[0.5, 1]$. We define SM2 as a function of a one-sided p-value such that the selection probability is lowest at $p = 0.5$ (i.e. effect $y = 0$). SM2 is asymmetric about $p = 0.5$ and has the general form

$$f_2(p) = \begin{cases} 1 & \text{if } 0 \leq p < 0.005 \\ \exp(d_1 \times p) & \text{if } 0.005 \leq p < 0.2 \\ \exp(d_2 \times p) & \text{if } 0.2 \leq p < 0.5 \\ \exp(d_3 \times (1-p)) & \text{if } 0.5 \leq p < 0.8 \\ d_4 & \text{if } 0.8 \leq p < 0.975 \\ d_5 & \text{if } 0.975 \leq p \leq 1 \end{cases} \tag{3.33}$$

where $d = (d_1, d_2, d_3, d_4, d_5) = (-0.5, -1, -2, 0.7, 0.9)$ for f_{2M} and $d = (-2, -4, -5, 0.4, 0.6)$ for f_{2E}. SM2 has the same form as SM1 on $[0, 0.5)$, and then has exponential increase on $[0.5, 0.8)$ and constant selection probabilities on $[0.8, 0.975)$ and $0.975, 1]$. SM1 and SM2 are shown in the top and bottom panels of Figure 3.1, respectively.

3.2.2 Simulation results

We generated $L = 200$ replicates for all 32 scenarios. RMSE for each scenario is shown in Figures 3.2 and 3.3, where Figure 3.2 is for SM1 and Figure 3.3 is for SM2. Left/right panels represent $\theta = 0.1$ or 0.5, respectively, and top/bottom panels represent extreme or moderate selection. The standard random effects meta-analysis is shown in blue, RoBMA is red, and stacking is green. The grey lines show the 10 selection models used for stacking. Stacking tends to have lower RMSE than both the standard model and RoBMA as sample sizes increase, and performs particularly well with extreme selection and small θ. Figures 3.4 and 3.5 show 95% interval coverage for every scenario. CI coverage rates for the standard model and RoBMA decrease considerably as sample sizes increase, while stacking maintains coverage near the 95% nominal level in all scenarios except a) extreme selection with small true mean θ, and b) moderate selection with small θ *and* average sample size of 80. In (a), no model shows coverage probabilities near the nominal level, although stacking is among the closest for each sample size. For (b), all models except one-sided selection with steps at $p = 0.025, 0.5$ have coverage probabilities drop below 0.9.

Figures 3.6 and 3.7 show bias for each model and simulation scenario. Stacking shows low bias for each sample size and combination of θ and selection severity, often having smaller bias than any individual model. One exception is the case of small

sample sizes, moderate selection, and large θ, where stacking tends to give weight to one-sided models that over-correct for publication bias. The standard model without any correction for publication bias has the largest bias in almost every scenario. The scenario with selection function 1, extreme selection, and $\theta = 0.1$ yields the largest bias across the board, with no model able to come close to reproducing the true mean.

3.2.3 Secondary simulation including sloped selection models

In a smaller simulation we additionally include 6 sloped selection models defined in Section 3.1.5 in the ensemble, giving $K = 17$ models to stack over, including the 6 sloped models, two Copas models, 8 stepped selection models, and the standard model. We use selection function f_{2E} and sample sizes $S = 10$ and 40. We compare results from the stacked model with the standard model and RoBMA with bias, 95% CI coverage, and RMSE.

Figure 3.8 shows bias, 95% CI coverage, and RMSE for each model. Stacking has smaller bias than either the standard model or RoBMA, and the difference is significant with an average sample size of $S = 40$ (stacking bias = 0.0007, RoBMA bias = 0.035, standard bias = 0.069). RMSE is comparable for RoBMA and stacking, around 0.11 for $S = 10$ and 0.06 for $S = 40$, and both are significantly better than the standard model. The standard model and RoBMA have 95% CI coverage less than the nominal level for $S = 40$ (RoBMA coverage = 0.88, standard coverage = 0.725) while stacking maintains the nominal coverage level.

Including the sloped selection models in stacking results in smaller absolute biases and smaller RMSE compared with the stacked model without sloped selection

models. Without including the sloped models, bias for the stacking model was -0.071 for $S = 10$ and -0.01 for $S = 40$, compared with -0.029 and 0.0007 when including the sloped models. RMSE without the sloped models was 0.137 or 0.060 for $S = 10$ or $S = 40$, and with the sloped models RMSE was 0.116 or 0.055 for $S = 10$ or $S = 40$.

3.3 Data analyses

We illustrate Bayesian stacking of selection models for publication bias on three datasets previously analyzed in the meta-analysis literature. In each example the stacked model is comprised of 17 models including Two-sidd (1) - (4), One-side (1) - (4), Slope-T (1) - (3), Slope-O (1) - (3), Mavridis Copas, Bai Copas, and the standard random effects model. We compare the stacked posterior distribution of the mean parameter θ with the posterior for θ from the standard meta-analysis model. In each example the results from the standard model are nearly or exactly equivalent to results reported in the original analyses, while the stacked model has posterior mean closer to the null value $\theta = 0$ and wider 95% CIs.

3.3.1 Second-hand smoke and lung cancer

Hackshaw et al. (1997) analyzed a set of 37 studies measuring the effects of second-hand smoke on the likelihood of developing lung cancer. The studies included in the analysis each measured the relative risk of lung cancer for women living with spouses who smoked vs women living with spouses who do not smoke. The authors performed a standard random-effects meta-analysis, finding a pooled relative risk (RR) of 1.24 (95% CI 1.13-1.36). The dataset from Hackshaw et al. (1997) has been

used to illustrate publication bias models in the past (Sterne et al., 2005; Ning et al., 2017; Takagi et al., 2006).

Model	Mean	SD	2.5%	97.5%	Stacking Weight
Stacked	0.111	0.076	-0.046	0.248	–
Standard	0.219	0.052	0.122	0.327	0.000
Mavridis	0.103	0.073	-0.036	0.255	0.244
Bai	0.167	0.081	-0.016	0.309	0.000
Two-side (1)	0.189	0.052	0.094	0.297	0.000
Two-side (2)	0.178	0.049	0.087	0.279	0.000
Two-side (3)	0.188	0.051	0.098	0.296	0.000
Two-side (4)	0.168	0.050	0.078	0.274	0.000
One-side (1)	0.179	0.055	0.072	0.293	0.000
One-side (2)	0.107	0.082	-0.064	0.249	0.266
One-side (3)	0.183	0.052	0.084	0.289	0.000
One-side (4)	0.130	0.059	0.019	0.248	0.342
Slope-T (1)	0.189	0.052	0.093	0.294	0.000
Slope-T (2)	0.183	0.050	0.093	0.285	0.000
Slope-T (3)	0.191	0.051	0.098	0.299	0.000
Slope-O (1)	0.101	0.091	-0.106	0.258	0.148
Slope-O (2)	0.156	0.062	0.034	0.276	0.000
Slope-O (3)	0.186	0.052	0.088	0.291	0.000

Table 3.1: Posterior summaries for each model using the second-hand smoke data from Hackshaw et al. (1997). The stacked model has a drastically different posterior distribution for θ than the standard meta-analysis, with a mean closer to 0 and larger SD. Models contributing to the stacked posterior are the Mavridis Copas model, one-sided stepped models with steps at (.025, .5) and (.025, .1), and a one-sided sloped selection model with knots at (.025, 5).

The main endpoint θ is the log-relative risk (log(RR)). Summaries for θ under each model are shown in Table 3.1. Four models contributed to the stacked posterior distribution: the Mavridis Copas model (weight = 0.244), one-sided stepped selection model with steps at $p = 0.025, 0.5$ (weight = .266), and the one-sided stepped selection model with steps at $p = 0.025, 0.1$ (weight = .342), and one-sided sliding

selection with knots at $p = .025, .5$. Figure 3.9 shows posterior distributions for θ for each model. The stacked mean of θ is 0.111, with a 95% CI of (-0.046, 0.248), which transforms to a mean relative risk of 1.12 (0.96, 1.28). Compared to the original results from Hackshaw et al. (1997), the stacked model estimates a mean increased risk of 12% rather than the 24% originally reported, with a much wider range of plausible values including the null value of 1.

3.3.2 Gender effects in grant proposals

Bornmann et al. (2007) compared the odds of a successful grant proposal for grants written by men compared to women using a dataset of 66 peer review procedures from 21 studies. Each study generally reported on one type of award and multiple peer review procedures for that award (e.g. TMR Marie Curie Fellowship for chemistry, engineering, mathematics, earth sciences, economics, physics and life sciences). Bornmann et al. (2007) fit a random effects meta-analysis model with the log-odds ratio (logOR) as the main endpoint θ. The empirical Bayes estimate for the logOR of grant acceptance for men compared to women was 0.07 (95% CI 0.01-0.13), indicating a significant effect in favor of men. A funnel plot of study-specific effect sizes against their standard errors shows potential evidence of publication bias, so we fit the stacking procedure to the set of 66 results.

Table 3.2 shows model-specific point estimates and 95% CIs. The two models yielding the highest stacking weights were the Mavridis Copas model (weight = .288), the Bai Copas model (weight = .475), the two-sided stepped selection model with steps at $p = 0.05, 0.50$ (weight = .236), and one-sided sloped selection with knots at $p = .025, .5$ (weight = .001). Figure 3.10 shows posterior distributions of

Model	Mean	SD	2.5%	97.5%	Stacking Weight
Stacked	0.051	0.034	-0.020	0.116	–
Standard	0.069	0.030	0.012	0.128	0.000
Mavridis	0.032	0.035	-0.038	0.104	0.288
Bai	0.060	0.032	-0.003	0.122	0.475
Two-side (1)	0.057	0.028	0.005	0.116	0.000
Two-side (2)	0.054	0.026	0.005	0.108	0.236
Two-side (3)	0.056	0.027	0.005	0.109	0.000
Two-side (4)	0.055	0.027	0.006	0.109	0.000
One-side (1)	0.055	0.031	-0.005	0.117	0.000
One-side (2)	0.012	0.041	-0.073	0.089	0.000
One-side (3)	0.049	0.031	-0.011	0.109	0.000
One-side (4)	0.041	0.032	-0.021	0.105	0.000
Slope-T (1)	0.061	0.028	0.008	0.117	0.000
Slope-T (2)	0.061	0.028	0.008	0.119	0.000
Slope-T (3)	0.057	0.027	0.006	0.113	0.000
Slope-O (1)	0.009	0.045	-0.084	0.090	0.001
Slope-O (2)	0.052	0.032	-0.011	0.114	0.000
Slope-O (3)	0.055	0.031	-0.004	0.118	0.000

Table 3.2: Posterior summaries for each model using the gender effects in grant proposals data from Bornmann et al. (2007). While the standard model yields a 95% CI excluding zero, the stacked posterior shifts the mean towards zero and the posterior CI includes zero. Models contributing to the stack are the Bai and Mavridis Copas models, two-sided stepped selection with change points at (.05, .5).

θ for the standard model, stacked model, and the three models contributing to the stacked model. The stacked posterior mean logOR was 0.051 with a 95% CI of (-0.020, 0.116), still indicating a trend of grant proposals favoring men, but the 95% interval includes the null value of 0.

3.3.3 Recidivism and cognitive behavioral therapy

Landenberger and Lipsey (2005) analyzed a collection of 58 studies measuring how

cognitive behavioral therapy (CBT) interventions are associated with recidivism in both adult and juvenile offenders. The authors fit a random effects meta-analysis model and report a mean odds ratio of 1.53 (logOR = 0.425) with $p < 0.001$. Figure 3.11 shows a funnel plot of the 58 studies. We that studies with larger standard errors tend to have larger logORs, indicating the possible presence of publication bias favoring studies with results that favor the CBT intervention.

Model	Mean	SD	2.5%	97.5%	Stacking Weight
Stacked	0.338	0.125	0.063	0.531	–
Standard	0.425	0.063	0.303	0.552	0.487
Mavridis	0.236	0.089	0.058	0.408	0.227
Bai	0.326	0.090	0.124	0.485	0.134
Two-side (1)	0.403	0.063	0.284	0.532	0.000
Two-side (2)	0.378	0.063	0.255	0.507	0.000
Two-side (3)	0.386	0.065	0.259	0.513	0.000
Two-side (4)	0.392	0.063	0.275	0.519	0.000
One-side (1)	0.396	0.066	0.267	0.527	0.000
One-side (2)	0.230	0.123	-0.047	0.438	0.028
One-side (3)	0.368	0.070	0.229	0.503	0.000
One-side (4)	0.368	0.071	0.223	0.506	0.000
Slope-T (1)	0.396	0.063	0.276	0.522	0.000
Slope-T (2)	0.397	0.064	0.276	0.526	0.000
Slope-T (3)	0.401	0.065	0.275	0.531	0.000
Slope-O (1)	0.214	0.131	-0.072	0.440	0.124
Slope-O (2)	0.387	0.069	0.249	0.522	0.000
Slope-O (3)	0.390	0.070	0.252	0.527	0.000

Table 3.3: Posterior summaries for each model from numerical example 3 using data from Landenberger and Lipsey (2005). The stacked model yields posterior distribution of θ with mean shifted towards zero and fatter tails compared with the standard model. Models contributing to the stack are the standard model, Mavridis and Bai Copas models, the one-sided stepped selection model with steps at (.025, .5), and the one-sided sloped selection model with knots at (.025, .5).

Table 3.3 shows posterior summaries for the mean logOR for each model. The

five models contributing to the stacked posterior were the standard model (weight = .487), Mavridis Copas (weight = 0.227), Bai Copas (weight = 0..134), one-sided stepped selection with change points at $p = 0.025, 0.5$ (weight = 0.028), and one-sided sloped selection with knots at $p = .025, .5$ (weight = .124). The stacked posterior mean logOR was 0.338 (95% CI 0.063, 0.531). Figure 3.12 shows posterior distributions for the standard model, stacked model, and the 3 models contributing to the stacked posterior. The stacked posterior logOR is shifted towards zero and is much more diffuse than the standard model.

3.4 Discussion

Our proposal to use Bayesian stacking of different selection models for publication bias is motivated as much by philosophy as it is by favorable statistical properties. As Rothstein et al. (2006) describes, publication bias is an umbrella term that encompasses a number of information suppression mechanisms, including language bias (favoring studies in English), availability bias (favoring easily available studies), cost bias (favoring studies that are low cost or free), familiarity bias (favoring studies from one's own discipline), and reporting bias (primary authors favoring results that tend toward statistical significance within a published article). We believe it is unlikely that any individual selection model captures the true mechanism of selecting studies for publication and inclusion in meta-analyses, so consideration of a number of different models is desirable. While Bayesian model averaging considers multiple models, it also assumes that one of the fitted models is the "true" model, which is undoubtedly false in this scenario. Bayesian stacking makes no such assumption, and instead weights models according to their predictive ability, which we see as a

philosophical advantage that also lends statistical advantages. Advantages include 1) the lack of a need to specify prior model probabilities, and 2) the ability to add multiple similar models without necessarily taking weight away from other models.

In our simulation we considered multiple designs for the number of studies per meta-analysis. Papers in the publication bias literature tend not to specify exactly what they mean by "number of studies" in a simulation. One option is to generate studies one at a time until a fixed target number of studies has survived the selection mechanism, in which case each simulation iteration for a given scenario will have the same number of studies included in the analysis. A second option is to start with a fixed initial number of studies representing the population of studies that has been conducted, and filtering the population through the selection mechanism. Here the number of studies per analysis would differ across simulation iterations for a given scenario, and will also differ in terms of the average number of studies across scenarios (e.g. a more severe selection mechanism will suppress more studies than a moderate selection mechanism). We chose to calculate the necessary number of initial studies for each simulation scenario such that the average number of studies across scenarios was 10, 20, 40, or 80 for small, medium, large, and very large meta-analyses respectively.

While stacking selection models performed well compared to individual selection models in simulations, especially in terms of bias and 95% interval coverage, none of the models, including the stacked model, performed well in the scenario with extreme selection bias and small true effect θ. The outlook in this particular scenario is grim, as it appears we do not have a model capable of fully mitigating the effects of publication bias even though selection models offer a preferable alternative to the standard model. Better results may be achieved by using stronger prior distributions

on parameters that determine the strength of selection (that is, ρ for Copas models, $\boldsymbol{\omega}$ for stepped selection models). Researchers may shy away from this practice, as it might induce over-correction of effect sizes in situations where publication bias is more moderate.

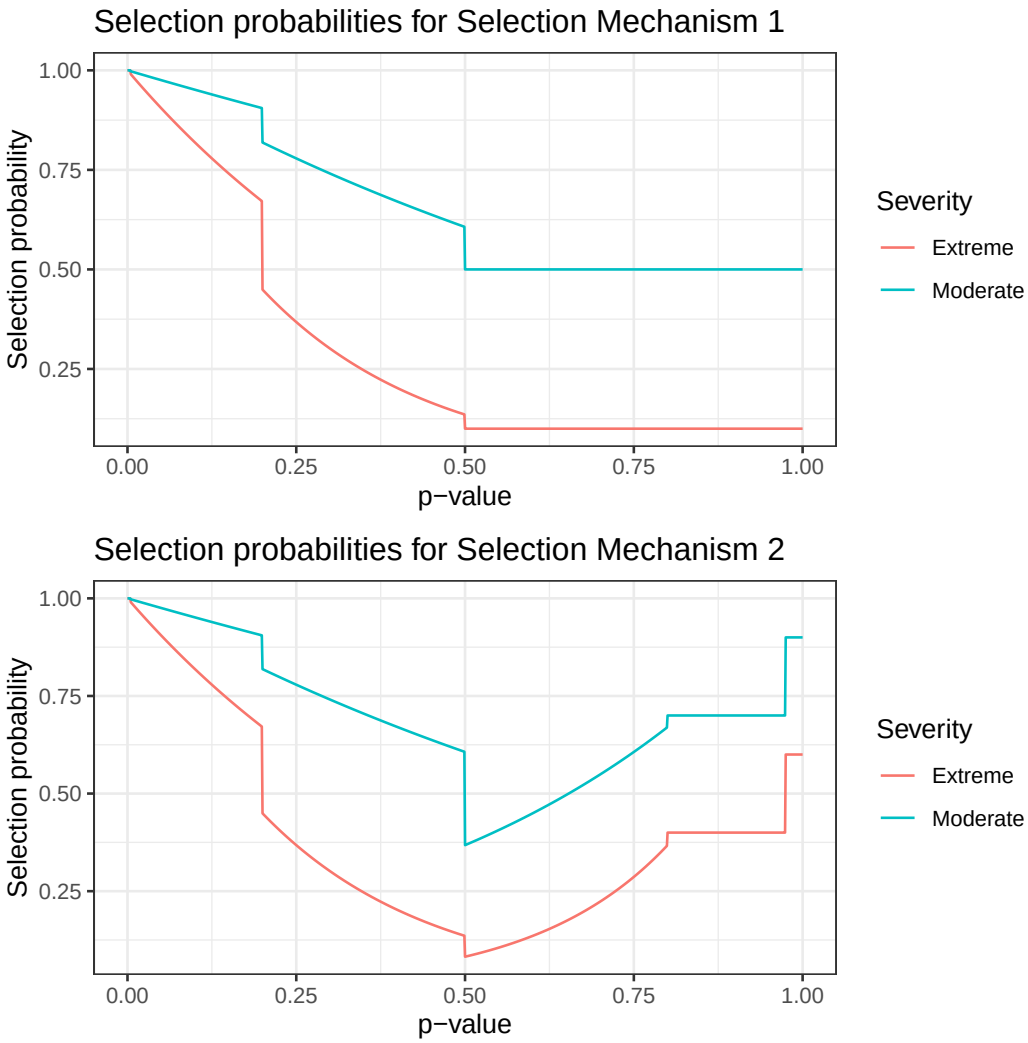

Figure 3.1: Selection Mechanisms (SMs) 1 and 2. SM1 has declining selection probabilities with increasing one-sided p-values, with change points at $p = 0.2$ and $p = 0.5$ and exponential decay between change points. SM2 has asymmetric two-sided selection with change points at $p = 0.005, 0.2, 0.5, 0.8$, and 0.975. The function minimum is at $p = 0.5$, i.e. two-sided p-value $= 1$.

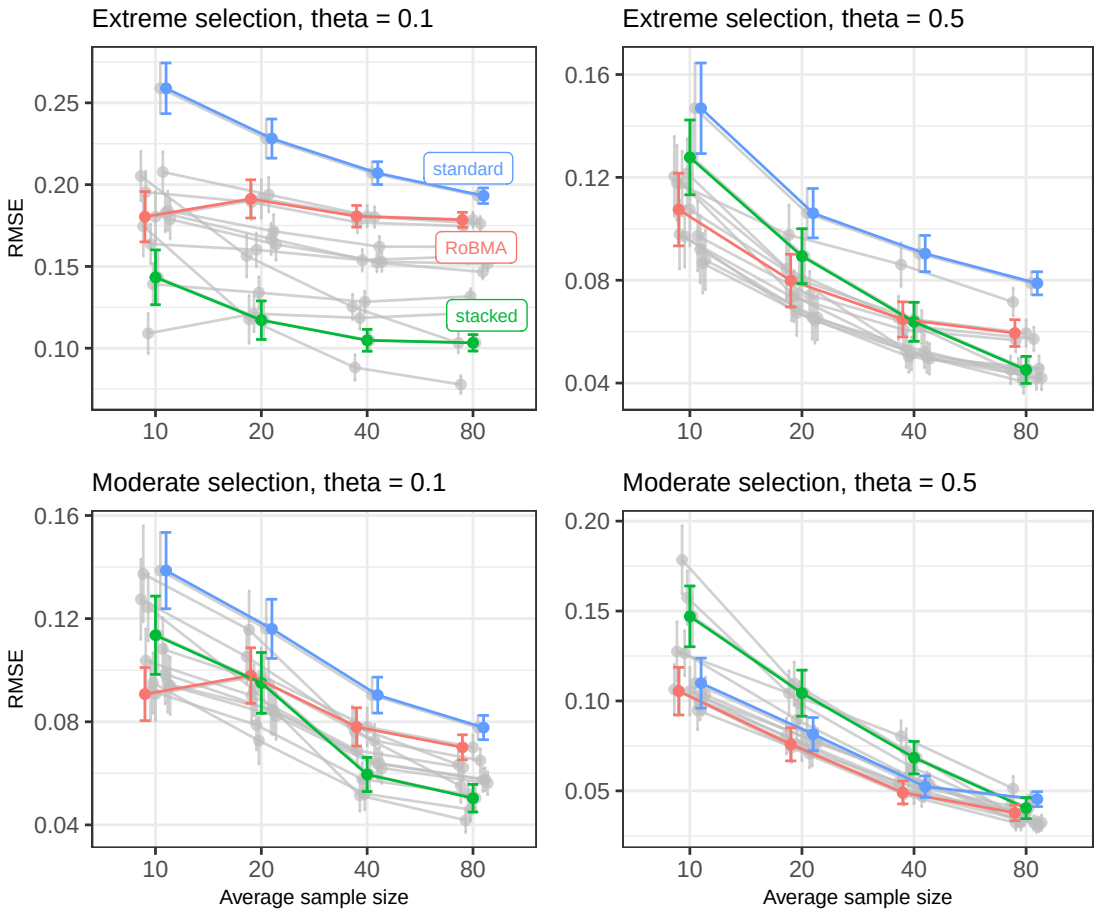

Figure 3.2: RMSE from 200 simulation replications using Selection Mechanism 1. Left and right panels have $\theta = 0.1$ and $\theta = 0.5$, respectively. Top and bottom panels have extreme or moderate selection severity respectively. The blue lines are for the standard random effects model, green is the stacked model, red is RoBMA, and grey lines are the individual selection models. Vertical error bars show $\pm 1.96 \times$ MCSE. The stacked model has much lower RMSE than the standard model or RoBMA with extreme selection and small θ, and with moderate selection and small θ when sample sizes are larger.

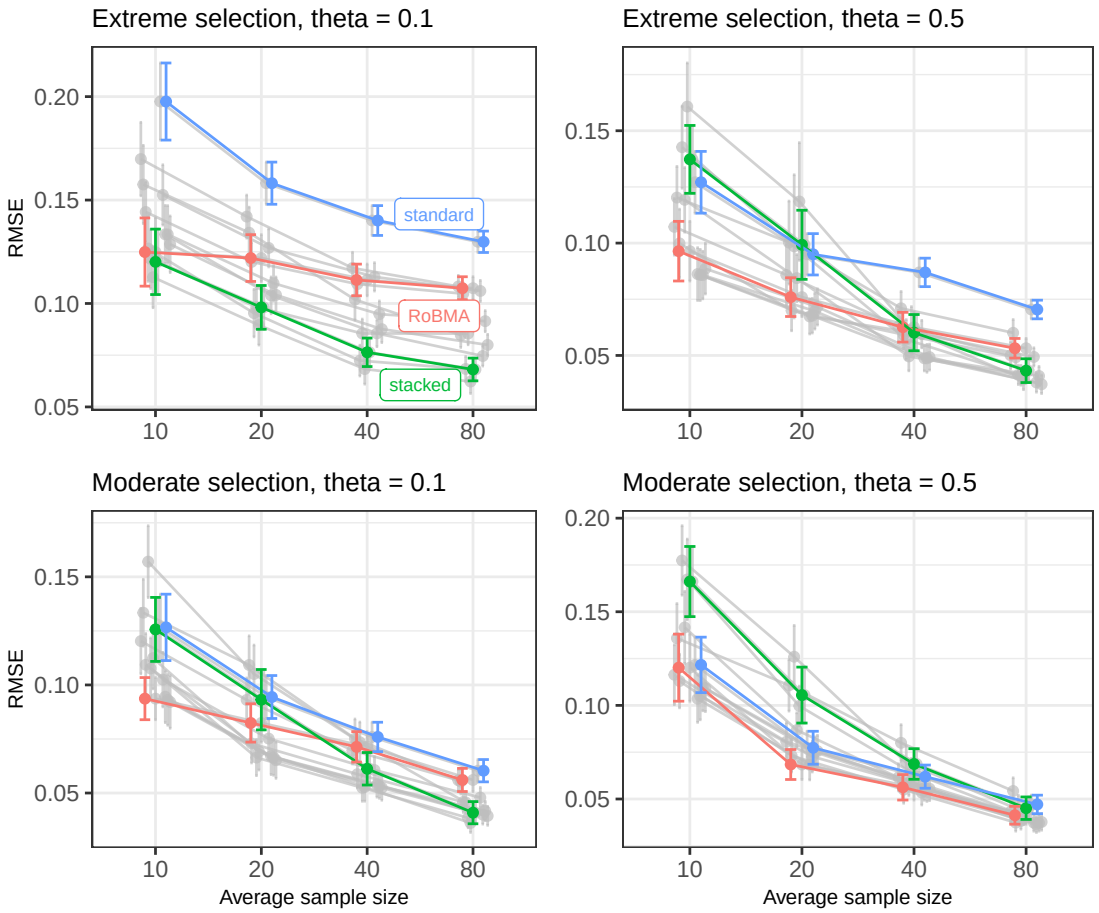

Figure 3.3: RMSE from 200 simulation replications using Selection Mechanism 2. Left and right panels have $\theta = 0.1$ and $\theta = 0.5$, respectively. Top and bottom panels have extreme or moderate selection severity respectively. The blue dots/lines are the standard random effects meta-analysis, green is the stacked model, red is RoBMA, and grey lines are the individual selection models. Vertical error bars show $\pm 1.96 \times$ MCSE. The stacked model has smaller RMSE than the standard model or RoBMA when there is extreme selection when $\theta = 0.1$ (upper left panel), but has higher RMSE with moderate selection and $\theta = 0.5$ (bottom right panel).

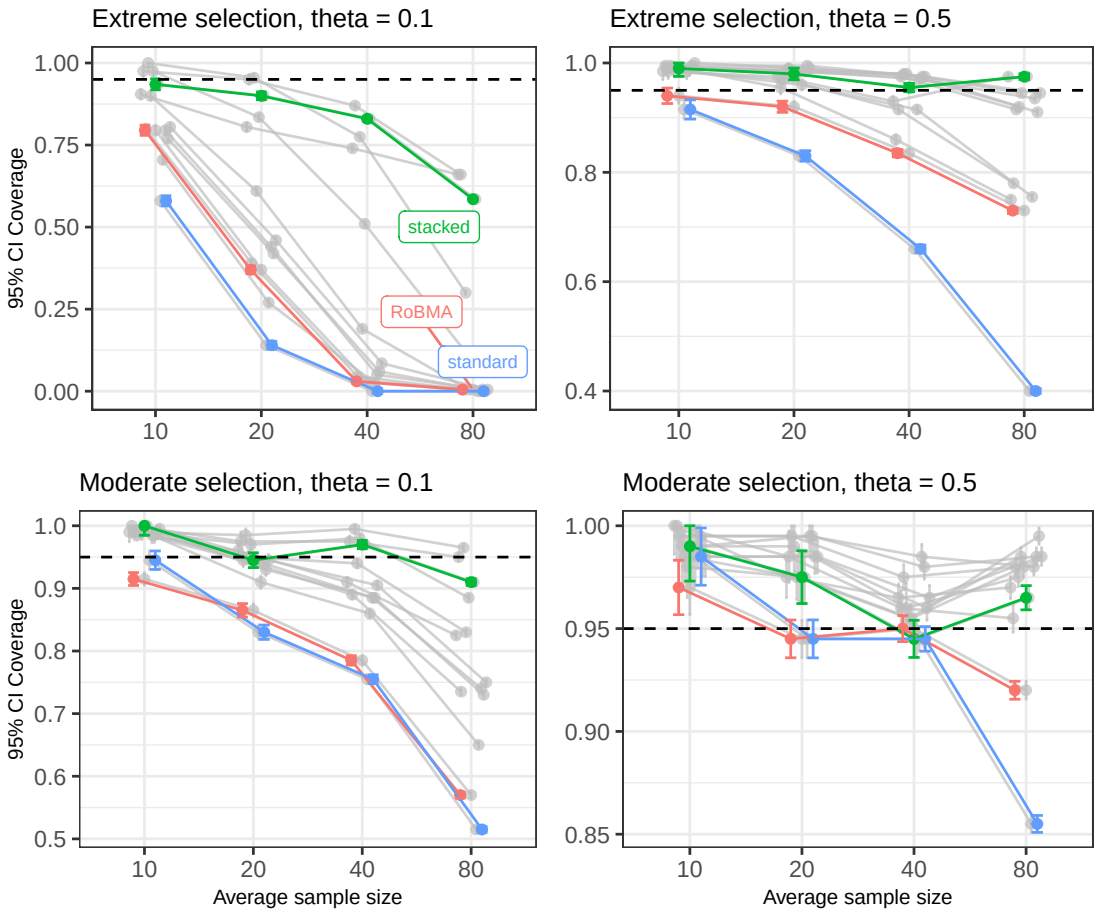

Figure 3.4: Proportion of 95% CIs covering the true mean θ from 200 simulation replications using Selection Mechanism 1. Left and right panels have $\theta = 0.1$ and $\theta = 0.5$, respectively. Top and bottom panels have extreme or moderate selection severity respectively. The blue line is the standard random effects model, green is the stacked model, red is RoBMA, and grey lines are the individual selection models. Vertical error bars show $\pm 1.96 \times$ MCSE. The stacked model has better or equal 95% CI coverage rates compared to the standard and RoBMA models for each combination of selection, θ, and sample size.

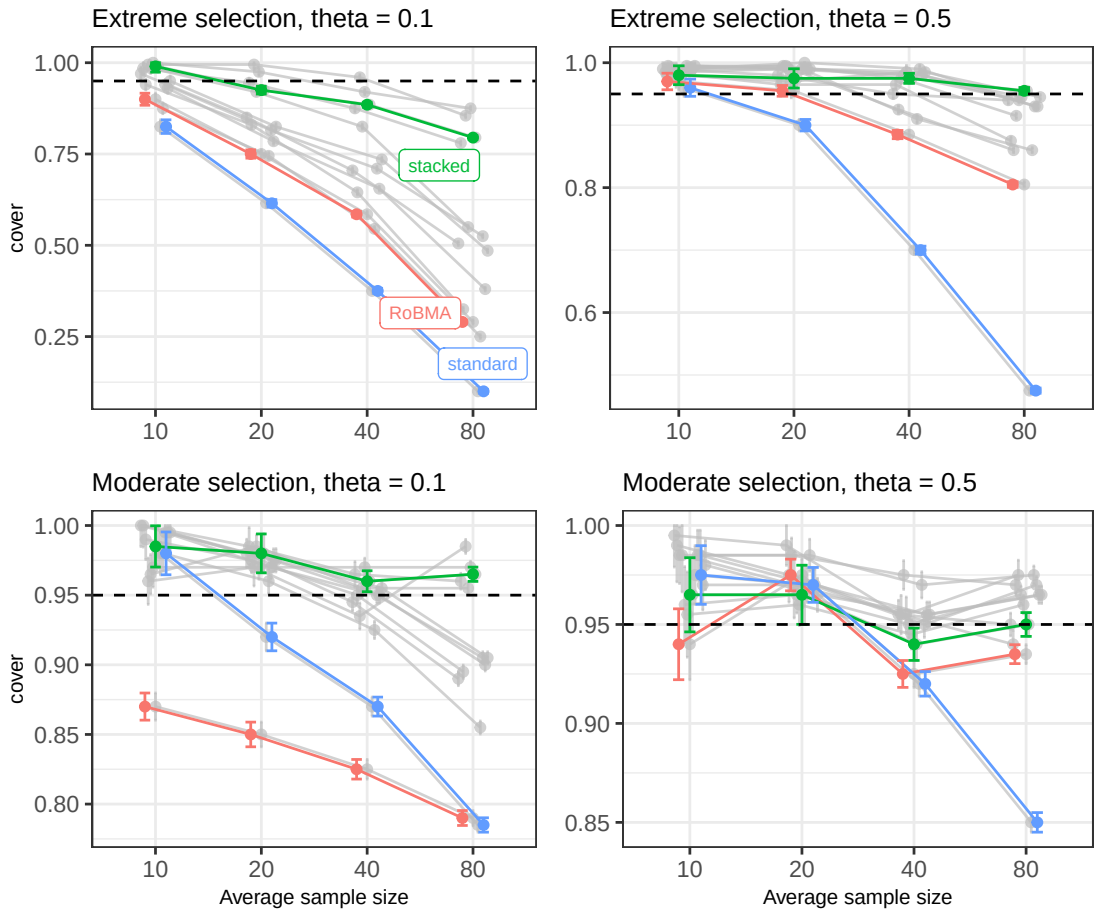

Figure 3.5: Proportion of 95% CIs covering the true mean θ from 200 simulation replications using Selection Mechanism 2. Left and right panels have $\theta = 0.1$ and $\theta = 0.5$, respectively. Top and bottom panels have extreme or moderate selection severity respectively. The blue line is the standard random effects model, green is the stacked model, red is RoBMA, and grey lines are the individual selection models. Vertical error bars show $\pm 1.96 \times$ MCSE. Both the standard model and RoBMA tend to see coverage fall well below the nominal 95% level as sample sizes increase, except with moderate selection and $\theta = 0.5$. Stacking either maintains at least the nominal 95% level or is among the closest models to 95% as sample sizes increase.

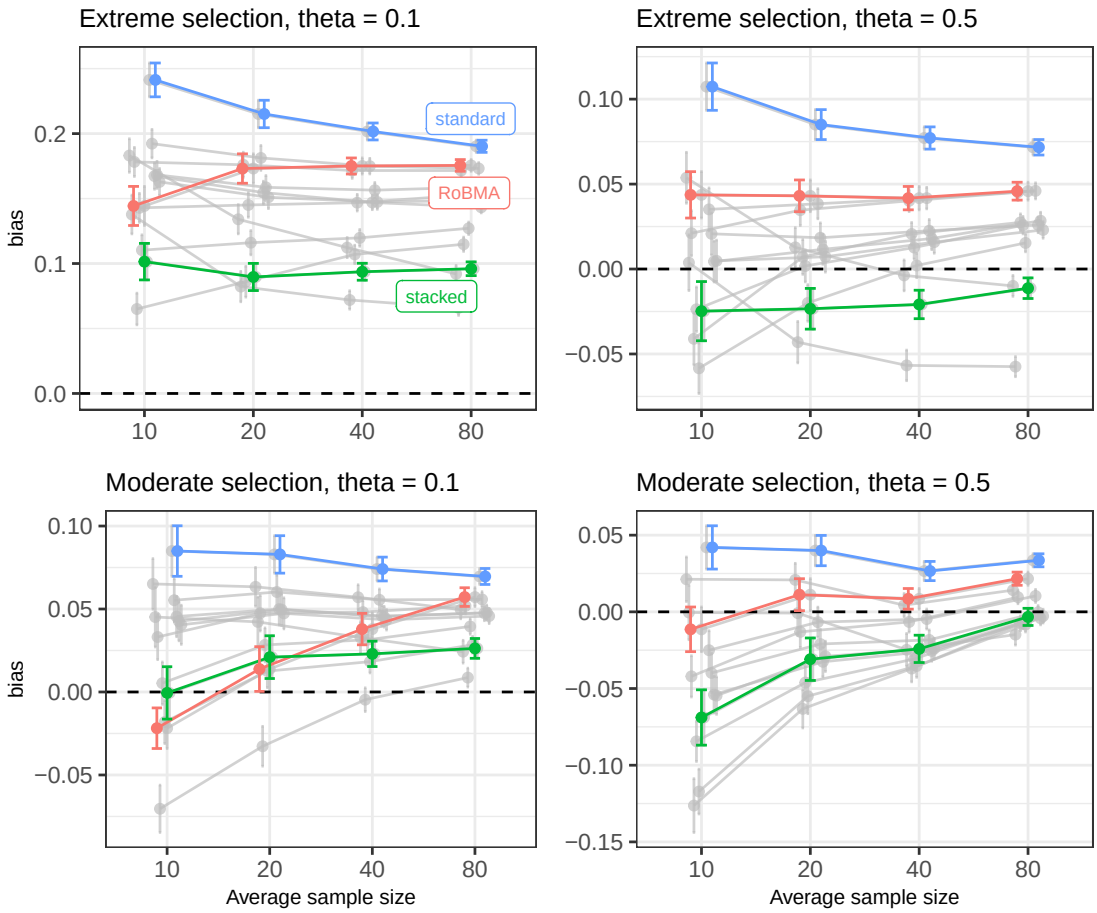

Figure 3.6: Bias from 200 simulation replications using Selection Mechanism 1. Left and right panels have $\theta = 0.1$ and $\theta = 0.5$, respectively. Top and bottom panels have extreme or moderate selection severity respectively. The blue dots/lines are the standard random effects meta-analysis, green is the stacked model, red is RoBMA, and grey lines are the individual selection models. Vertical error bars show $\pm 1.96 \times$ MCSE. The stacked model has smaller bias than the standard model and RoBMA regardless of sample size when there is extreme selection and also has smaller bias with moderate selection and larger sample sizes.

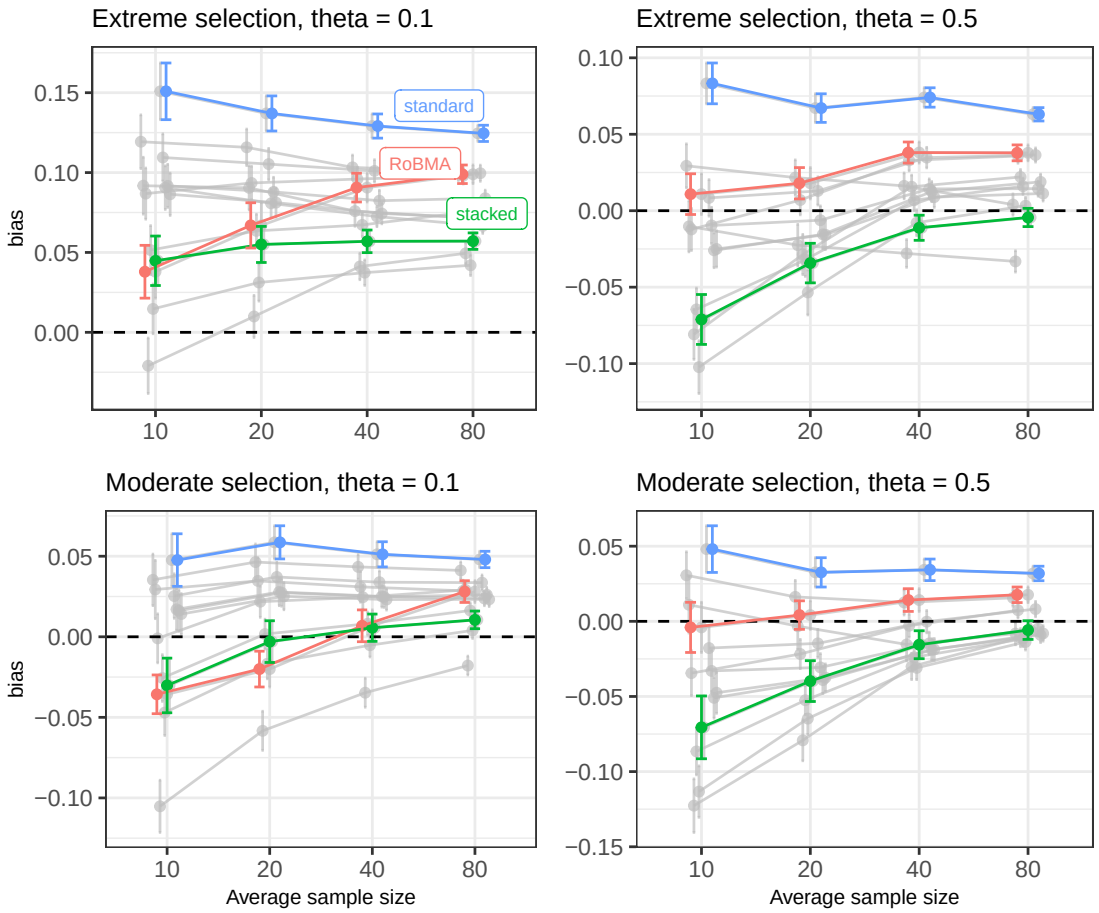

Figure 3.7: Bias from 200 simulation replications using Selection Mechanism 2. Left and right panels have $\theta = 0.1$ and $\theta = 0.5$, respectively. Top and bottom panels have extreme or moderate selection severity respectively. The blue dots/lines are the standard random effects meta-analysis, green is the stacked model, red is RoBMA, and grey lines are the individual selection models. Vertical error bars show $\pm 1.96 \times$ MCSE. The stacked model has the smallest bias in each panel when sample sizes are large (40 or 80 studies per analysis). RoBMA has small bias when $\theta = 0.5$ and sample sizes are smaller.

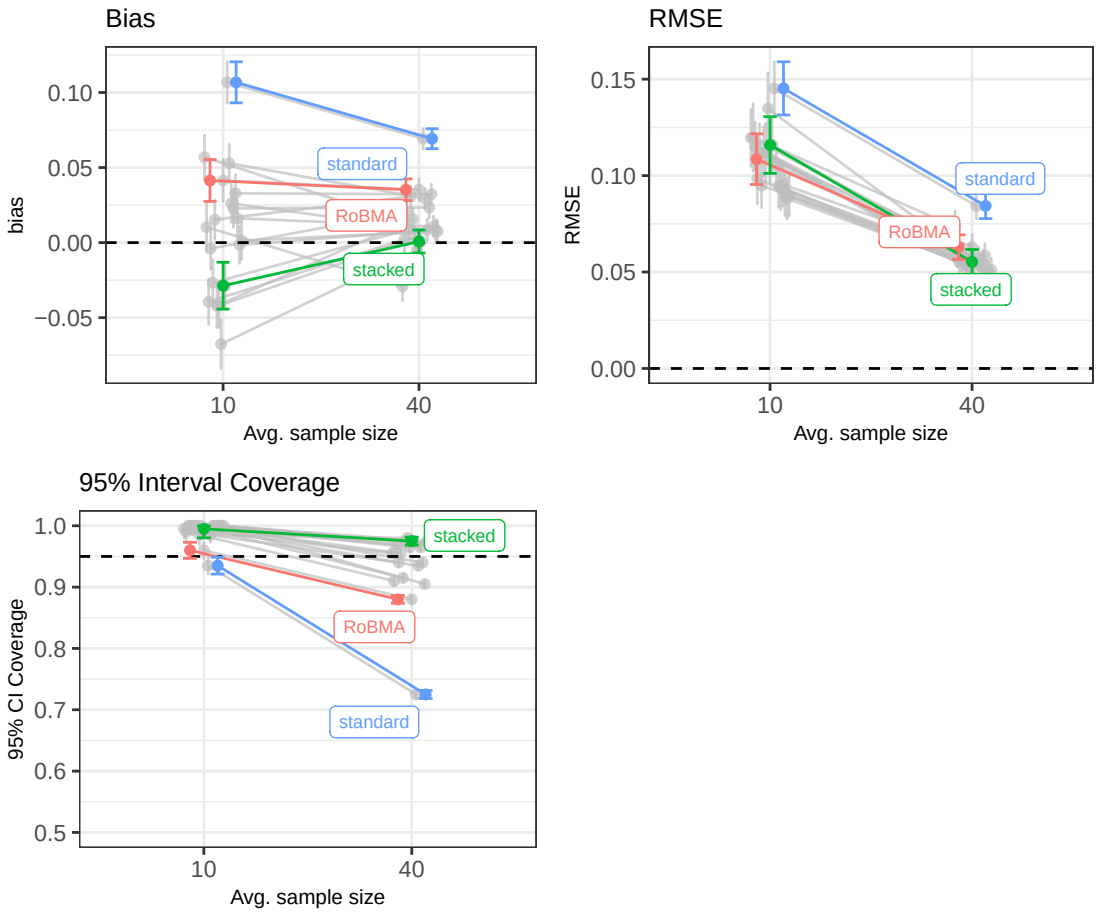

Figure 3.8: Bias, RMSE, and 95% CI coverage from 200 simulation iterations including sloped selection models. The standard random effects model is blue, RoBMA is red, and stacking is green, and selection models used for stacking are grey. Error bars represent $\pm 1.96 \times$ MCSE. Bias is smaller for stacking than for RoBMA or the standard model with average sample sizes of 10 and 40, but the difference in biases is only significant with 40 studies. RMSE is comparable for RoBMA and stacking, and both are better than the standard model. Only stacking maintains at least the 95% nominal coverage level as the number of studies increases.

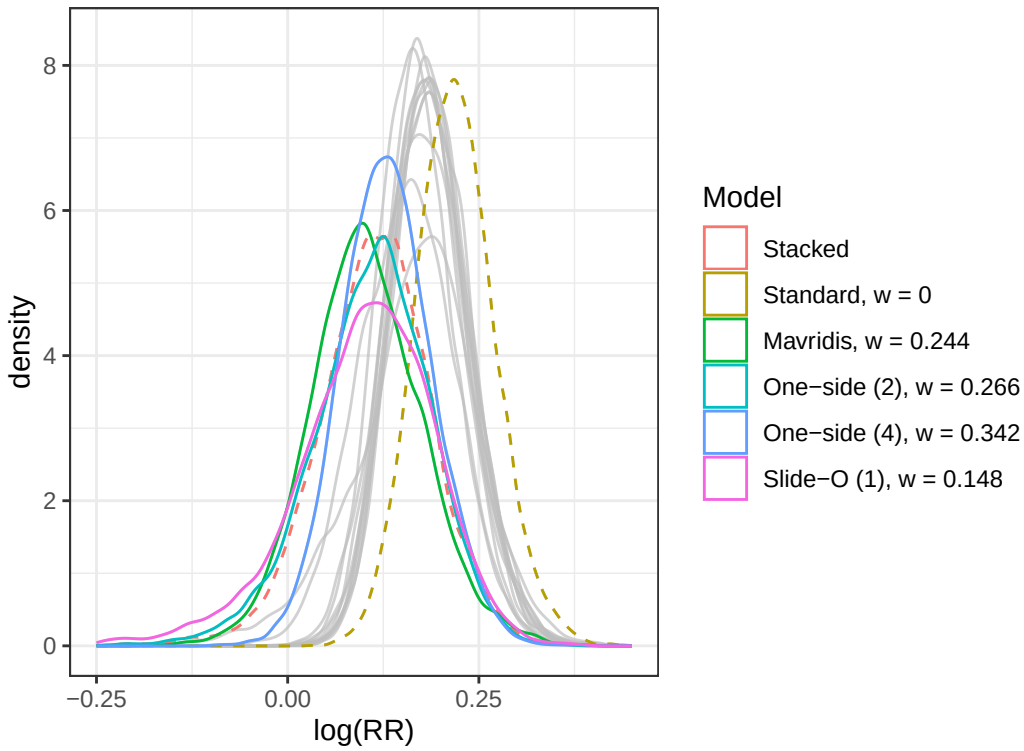

Figure 3.9: Posterior distributions of θ for each selection model using lung cancer data from Hackshaw et al. (1997). Colored lines show models where stacking weight was at least 0.01 (1%). Yellow dashed line is the standard meta-analysis. Red dashed line shows stacked posterior distribution.

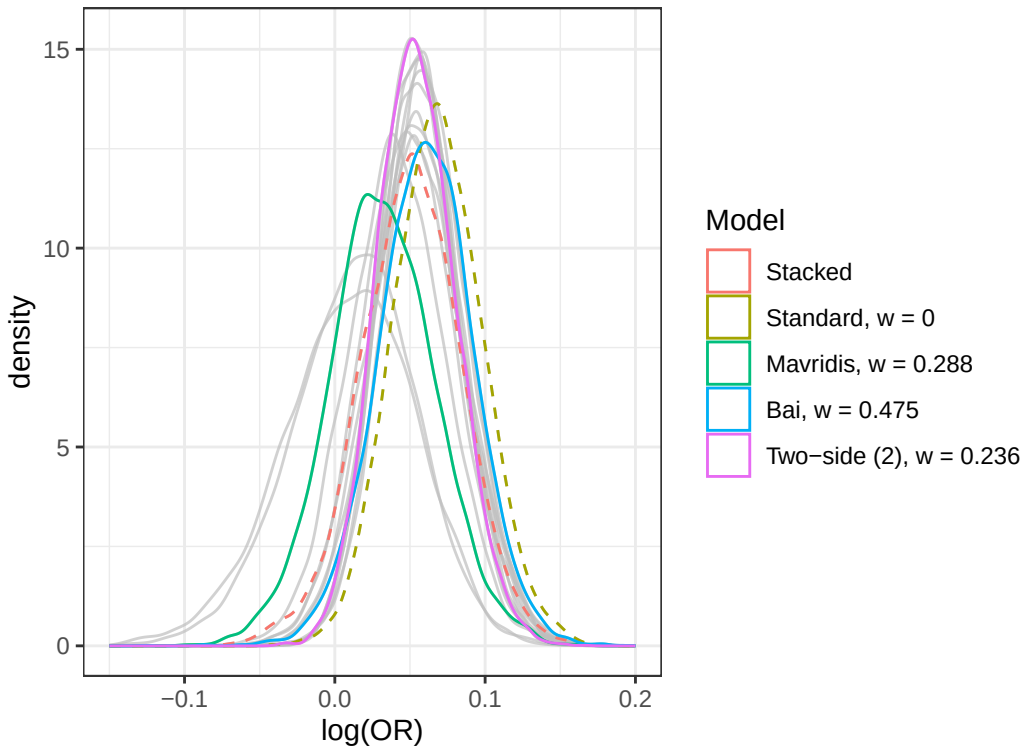

Figure 3.10: Posterior distributions of θ for each selection model using grant application data from Bornmann et al. (2007). Colored lines show models where stacking weight was at least 0.01 (1%). Yellow dashed line is the standard meta-analysis. Red dashed line shows stacked posterior distribution.

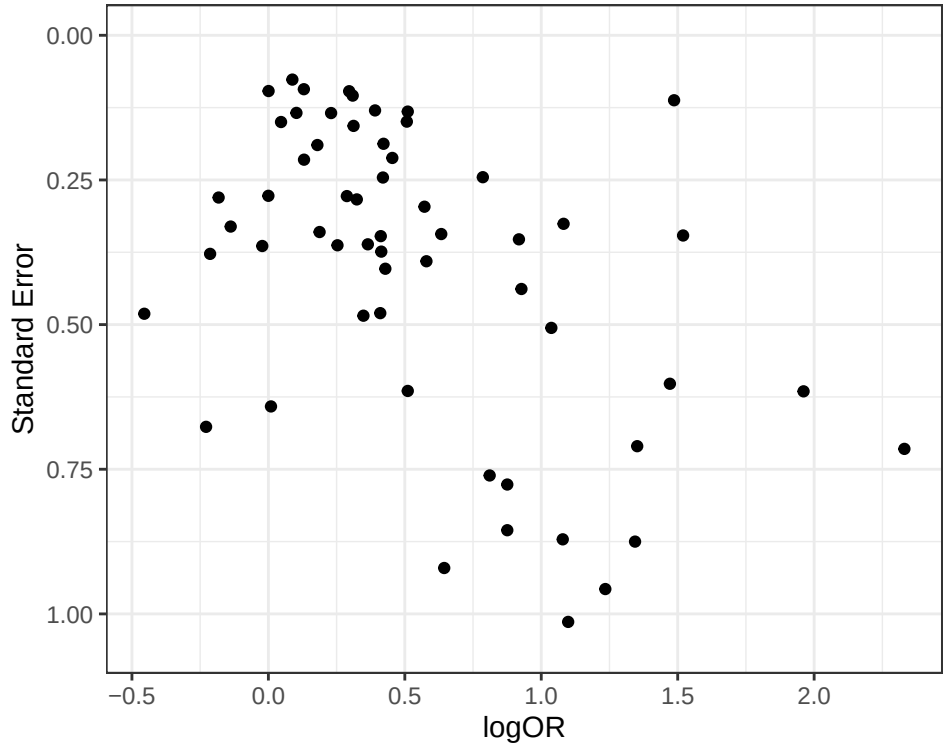

Figure 3.11: Inverted funnel plot of studies from Landenberger and Lipsey (2005), with log-odds ratios (logOR) on the x-axis and their standard errors on the y-axis. There is strong asymmetry, where studies with larger standard errors tend to have larger logORs, indicating the likely presence of publication bias.

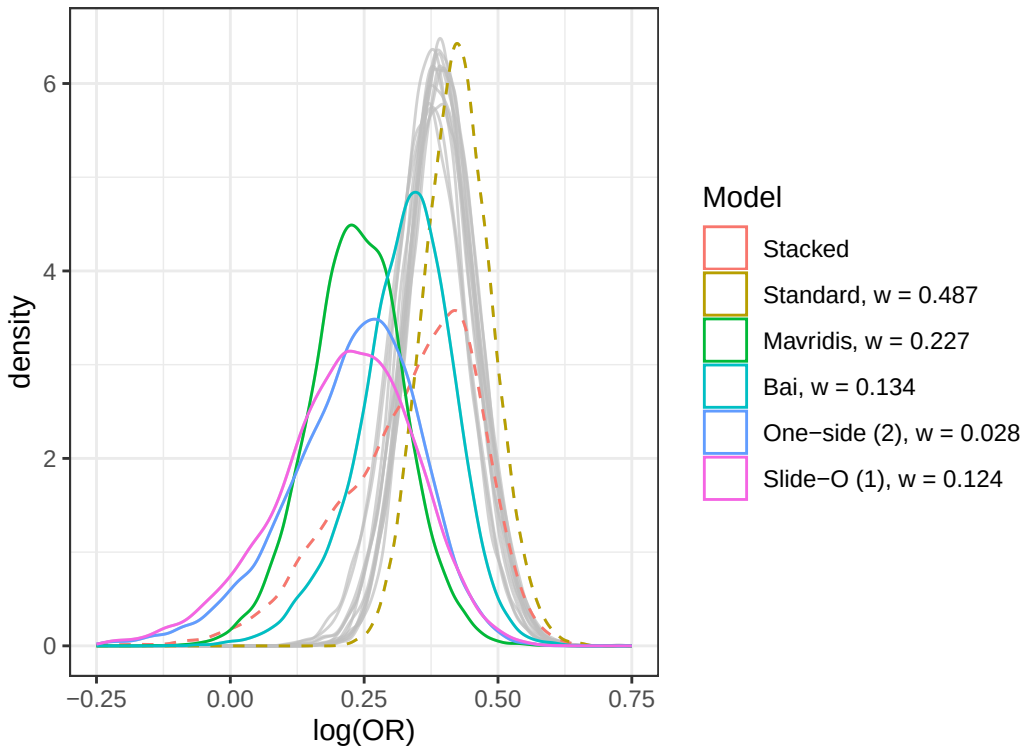

Figure 3.12: Posterior distributions of θ for each selection model using recidivism data from Landenberger and Lipsey (2005). Colored lines show models where stacking weight was at least 0.01 (1%).Yellow dashed line is the standard meta-analysis. Red dashed line shows stacked posterior distribution.

CHAPTER 4

Covariance Modeling in Meta-analysis with Regularized Horseshoe Priors

In meta-analysis, a usual aim is to estimate some parameter by averaging results across a number of studies that estimate some parameter. A fixed-effects model assumes that heterogeneity in results across studies is due to solely to sampling variation. A random-effects (RE) model instead assumes that study-specific parameters are drawn from a common distribution with some variance to be estimated. If there are two or more REs per study, the REs may either be assumed to be a priori independent from each other or they are jointly modeled as multivariate normal (MVN) or some other multivariate distribution. In a Bayesian meta-analysis with multiple REs one needs to assign prior distributions to both the vector of mean parameters and the covariance matrix $\mathbf{\Sigma}$. Priors for mean parameters are generally set to be diffuse normal distributions, but there is less consensus on appropriate priors distribution for the covariance matrix $\mathbf{\Sigma}$, where different choices may lead to different posterior inferences (Wang et al., 2020), especially when there are few studies in the analysis.

Observational 2×2 contingency table data arises when neither row- nor column-totals of the 2×2 table are fixed by study investigators. Bayesian meta-analysis

models for observational 2×2 contingency table data (Gibson et al., 2018; Ma et al., 2018) have three random effects (3RE) per study. Existing 3RE models either model the random effects with independent normal distributions with unknown means and variances (Gibson et al., 2018), or with a trivariate normal distribution with unknown vector of means and unknown covariance matrix $\mathbf{\Sigma}$ (Ma et al., 2018). Ma et al. (2018) gives $\mathbf{\Sigma}$ an inverse-Wishart (IW) prior, which is often used because it is conjugate in a multivariate normal model. The IW prior is inflexible because there is only one degree of freedom (df) parameter, ν, for all of the RE variances – the diagonal elements of $\mathbf{\Sigma}$ – and the IW prior gives little prior mass to values near zero for RE variances. If the true variance of one of the random effects is near zero, the observed correlations associated with that random effect are very noisy if there are few studies or the sample sizes per study are small, and the IW prior will have a very diffuse posterior distribution for correlations.

Network meta-analysis (NMA) extends traditional meta-analysis to the case of multiple treatment comparisons. The two main approaches to NMA are the contrast-based (CB) (Lu and Ades, 2004; Dias et al., 2013) and arm-based (AB) models (Zhang et al., 2014; Hong et al., 2016; Zhang et al., 2017). Let T be the number of treatments including some reference treatment and say the dichotomous outcome is presence or absence of an adverse event. The CB model treats log-odds ratios for each non-reference treatment relative to the reference treatment as exchangeable, whereas the AB model treats the log-odds of the event for each treatment, including the reference treatment, as exchangeable. The AB-NMA model has advantages over CB-NMA in estimating absolute risks (ARs) and functions of ARs, such as AR differences and marginal log-odds ratios. Each study in the AB-NMA typically provides data on some subset of the T treatments, often only on two. AB-NMA treats missing

treatments as missing at random (MAR), and models all treatments for each study as coming from a T-dimensional MVN distribution with $T \times T$ covariance matrix $\boldsymbol{\Sigma}$ and mean vector $\boldsymbol{\mu}$. A number of prior distributions have been proposed for modeling $\boldsymbol{\Sigma}$. The IW prior directly on $\boldsymbol{\Sigma}$ is among the most popular. Another approach uses a *separation strategy* that decomposes $\boldsymbol{\Sigma} = WPW$ where $T \times T$ diagonal matrix W has standard deviations σ_t, $t = 1, \ldots, T$, along the diagonal, and P is a correlation matrix with 1's on the diagonal and off-diagonal elements $P_{t,t'}$ which are the correlation between the random effects for treatments t and t' (Barnard et al., 2000). Priors are then placed on each σ_t and on P or elements of P. Covariance priors using the separation strategy have been shown to estimate ARs and mean treatment effects more accurately than the IW prior (Wang et al., 2020), but may still suffer from a lack of information in the data when certain treatments are only included in a few studies. The standard AB-NMA model includes only fixed treatment effects and random study-treatment effects, which induces positive correlations among treatment random effects if event rates tend to be higher or lower across the board for a given study. Thus, if variation in different treatments can be explained by a single study random effect, the true treatment RE variances may be close to zero and the IW and separation strategy priors may not provide adequate shrinkage.

Methods for inducing sparsity in a covariance matrix $\boldsymbol{\Sigma}$ or precision matrix $\boldsymbol{\Sigma}^{-1}$ have been proposed in both the frequentist and Bayesian covariance estimation literature. Dempster (1972) first proposed a method of setting certain elements of $\boldsymbol{\Sigma}^{-1}$ be exactly zero, and termed the method *covariance selection*. Friedman et al. (2008) introduced the graphical LASSO (GLASSO) to estimate a sparse $\boldsymbol{\Sigma}^{-1}$ using an L_1 penalty on the elements of $\boldsymbol{\Sigma}^{-1}$. Bayesian versions of the GLASSO have

since been proposed (Banerjee and Ghosal, 2015). Wong et al. (2003) and Cripps et al. (2005) instead decompose the precision matrix as $\mathbf{\Sigma}^{-1} = ABA$, where A is a diagonal matrix containing the square roots of partial precisions and B is a partial correlation matrix. Partial precision and partial correlation are the precision or correlation given the other parameters. Cripps et al. (2005) induce sparsity in $\mathbf{\Sigma}^{-1}$ by setting spike-and-slab priors on the partial correlations B_{ij}, where B_{ij} is the $(i,j)^{\text{th}}$ element of B. Chen and Dunson (2003) and Cai and Dunson (2006) decompose $\mathbf{\Sigma}$ as $\mathbf{\Sigma} = \Omega\Gamma\Gamma'\Omega$ where Ω is a diagonal matrix with elements ω_i that are proportional to random effects standard deviations and Γ is lower triangular with 1's on the diagonal and off-diagonal elements related to random effects correlations. They place spike-and-slab priors on the elements of Ω and Γ to allow random effect variances and correlations to be exactly zero.

We propose a new covariance selection prior for the 3RE and AB-NMA models which uses the decomposition $\mathbf{\Sigma} = \Omega\Gamma\Gamma'\Omega$ and the regularized horseshoe (RHS) prior (Piironen and Vehtari, 2017) on elements of Ω. The RHS prior allows RE variances to shrink to nearly zero which allows REs to effectively drop from the model. We introduce a new *conditional shrinkage prior* for elements γ_{ij} of Γ that can regularize correlations conditional on RE variances, and we develop a simple method for setting prior parameters for RHS prior for both 3RE and AB-NMA models. Using synthetic and real data examples we compare the new RHS prior with regularly used default priors using expected log-predictive density (elpd) and by comparing posterior distributions of quantities of interest for the 3RE and AB-NMA models. To distinguish between study-level variation in event rates and study-treatment variation we offer a new formulation for the AB-NMA model that models study main effects and study-treatment random effects.

Section 4.1 describes 3RE and AB-NMA meta-analysis models and details a new AB-NMA model formulation for separating random study effect variance from random study-treatment variances and correlations. We describe the new covariance selection prior in Section 4.2 and describe how to choose default input values and incorporate prior information. We illustrate the covariance selection prior in Section 4.3 using both synthetic and real data examples. The paper closes with discussion.

4.1 Meta-analysis models

We describe two meta-analysis models, the 3RE model for observational contingency table data and arm-based network meta-analysis (AB-NMA) model for MTC data. The 3RE and AB-NMA models have multiple random effects which can be modeled as multivariate normal.

4.1.1 3RE meta-analysis model for observational contingency table data

In a meta-analysis of observational contingency table data, each study $i = 1, \ldots, S$ reports a 2×2 table of counts n_{ijk} of people with rows $j = 0, 1$ defined by the absence ($j = 0$) or presence ($j = 1$) of a risk factor (RF), denoted $\overline{\text{RF}}$ and RF, and columns $k = 0, 1$ defined by no adverse event ($\overline{\text{E}}$, $k = 0$) or adverse event (E, $k = 1$). Let $n_{i1} = n_{i10} + n_{i11}$ and $n_{i0} = n_{i00} + n_{i01}$ be the number of people in study i with or without the risk factor, respectively, and $N_i = n_{i1} + n_{i0}$ be the total sample size in study i. Let π_{ij} be the unknown probability of an adverse event for a patient in study i, group j, and ψ_i be the unknown prevalence of the risk factor in the population studied by study i. A three-random effect (3RE) model for observational

2×2 contingency table data is

$$n_{ij1}|\pi_{ij} \sim \text{Bin}(n_{ij}, \pi_{ij}) \tag{4.1}$$

$$\text{logit}(\pi_{ij}) = \begin{cases} \beta_i - \frac{\delta_i}{2} & j = 0 \\ \beta_i + \frac{\delta_i}{2} & j = 1, \end{cases} \tag{4.2}$$

$$n_{i1}|\psi_i \sim \text{Bin}(N_i, \psi_i), \tag{4.3}$$

where β_i is a random study effect for the average log-odds of the event between groups $j = 0$ and $j = 1$, and δ_i is a random study effect for the log-odds ratio of the event. The random effects β_i, δ_i, and $\nu_i = \text{logit}(\psi_i)$ can be modeled with independent normal distributions if there is no suspected correlation between them, or with a multivariate normal distribution if there are suspected correlations. For example, consider the scenario where we believe the probability of an event for those without the risk factor is relatively constant across studies (i.e. $\pi_{i0} \approx \pi_0$ for all i), but that the event rate π_{i1} for those with the risk factor varies across studies. This would imply a positive correlation between log-odds ratios δ_i and the average log-odds of an event β_i. If there are suspected correlations between the random effects, we give them a multivariate normal prior distribution with mean parameter $(\beta_0, \delta_0, \nu_0)$ and covariance matrix $\mathbf{\Sigma}$

$$\begin{pmatrix} \beta_i \\ \delta_i \\ \nu_i \end{pmatrix} |\Theta \sim \text{N}\left(\begin{pmatrix} \beta_0 \\ \delta_0 \\ \nu_0 \end{pmatrix}, \mathbf{\Sigma} = \begin{pmatrix} \sigma_\beta^2 & \rho_{\beta\delta}\sigma_\beta\sigma_\delta & \rho_{\beta\nu}\sigma_\beta\sigma_\nu \\ \rho_{\beta\delta}\sigma_\beta\sigma_\delta & \sigma_\delta^2 & \rho_{\delta\nu}\sigma_\delta\sigma_\nu \\ \rho_{\beta\nu}\sigma_\beta\sigma_\nu & \rho_{\delta\nu}\sigma_\delta\sigma_\nu & \sigma_\nu^2 \end{pmatrix} \right), \tag{4.4}$$

where $\Theta = (\beta_0, \delta_0, \nu_0, \sigma_\beta^2, \sigma_\delta^2, \sigma_\nu^2, \rho_{\beta\delta}, \rho_{\beta\nu}, \rho_{\delta\nu})$. A common prior choice for $\mathbf{\Sigma}$ is the inverse-Wishart $\text{IW}_M(V, v)$ with degrees of freedom v and scale matrix V where the

prior $\mathbf{\Sigma}_{M\times M}|V, v \sim \text{IW}_M(V, v)$ has density

$$\text{IW}_M(\mathbf{\Sigma}; V, v) = \frac{|V|^{v/2}}{2^{vM/2}\Gamma_M(v/2)}|\mathbf{\Sigma}|^{-(v+M+1/2}e^{-\frac{1}{2}\text{tr}(V\mathbf{\Sigma}^{-1})} \tag{4.5}$$

where M is the dimension of $\mathbf{\Sigma}$, Γ_M is the multivariate gamma function, and $\text{tr}(\cdot)$ is the trace function. With scale matrix V and degrees of freedom v, the prior mean for $\mathbf{\Sigma}$ is $V/(M - v - 1)$ for $v > M - 1$. The IW prior is popular because it is conjugate in a multivariate normal model. However, the IW prior is inflexible in that the prior uncertainty for each variance component is controlled by a single degrees of freedom parameter and there is very little prior weight given to values near zero. We instead propose the covariance selection approach to be presented in Section 4.2 to show our prior belief that certain random effect variances may be near zero.

4.1.2 Arm-based network meta analysis (AB-NMA)

Multiple treatment comparisons (MTC) data has S studies indexed by $i = 1, \ldots, S$, and a set of T treatments $\mathcal{T} = \{1, \ldots, T\}$ indexed by $t = 1, \ldots, T$. Each study i reports the number of events y_{it} and number of subjects n_{it} for some subset of treatments $\mathcal{T}_i \in \mathcal{T}$. An arm-based network meta-analysis (AB-NMA) (Hong et al., 2016) for MTC data is given by the model

$$y_{it}|p_{it} \sim \text{Bin}(n_{it}, p_{it}) \tag{4.6}$$

$$\text{logit}(p_{it}) = \mu_t + \eta_{it}, \tag{4.7}$$

where p_{it} is the unknown probability of an event for the t^{th} treatment in the i^{th} study, μ_t is the mean log-odds of the event for the t^{th} treatment, and η_{it} is a study random

effect for the t^{th} treatment. We call linear model (4.7) the standard model.

As each study only reports on a subset of the T treatments, the AB-NMA model treats unreported treatment arms as missing data. The mean effects μ_t are given vague independent normal prior distributions with known variance s^2_μ, and the random effects η_{it} are given a multivariate normal distribution

$$\mu_t | s^2_\mu \sim \mathrm{N}(0, s^2_\mu) \tag{4.8}$$

$$(\eta_{i1}, \ldots, \eta_{iT})' | \boldsymbol{\Sigma} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{4.9}$$

where $\boldsymbol{\Sigma}$ is a $T \times T$ unstructured covariance matrix of random effects variances for each treatment on the diagonal and the covariances between treatment random effects on the off-diagonal.

We have seen in our work that often there are very high observed correlations between random effects. This is likely induced by a main study effect η_{i0} that applies to all treatments for a given study, and additional variation apart from the main study effect may be very small. An alternative parameterization to the standard model (4.6) - (4.9) adds a main study effect η_{i0} to the linear predictor

$$\begin{aligned} \mathrm{logit}(p_{it}) &= \mu_t + \eta_{i0} + \eta_{it} \\ \eta_{i0} | \sigma^2_\eta &\sim \mathrm{N}(0, \sigma^2_\eta) \\ \eta_{it} | \sigma^2_t &\sim \mathrm{N}(0, \sigma^2_t), \end{aligned} \tag{4.10}$$

where random effects η_{it} and $\eta_{it'}$ are independent. We call linear model (4.10) the separate variance (SV) model.

Several priors for $\boldsymbol{\Sigma}$ under model (4.6) - (4.9) are summarized in Wang et al.

(2020). The most common prior is the conjugate $\text{IW}_T(V, T+1)$ with degrees of freedom $T+1$ and where V is a known $T \times T$ scale matrix (Hong et al., 2016; Zhang et al., 2017). Another option is the separation strategy $\mathbf{\Sigma} = WPW$, where W is a diagonal matrix containing random effect standard deviations $(\sigma_1, \ldots, \sigma_T)'$ and P is a $T \times T$ correlation matrix (Barnard et al., 2000). We then place independent priors on $\sigma_1, \ldots, \sigma_T$ and P, with several options available for both. The most common priors for standard deviations σ_t are Uniform(0, 5), half-normal, or half-t. Meta-analysts often opt for a compound-symmetric structure for P with 1 on the diagonal and all correlations $\rho_{tt'} = \rho$, where $\rho \sim \text{Uniform}(-\frac{1}{T-1}, 1)$ as a vague prior that ensures P is positive definite. If P is unstructured, another popular prior is the $\text{LKJ}(a_P)$ prior for correlation matrices (Lewandowski et al., 2009), where the shape parameter a_P determines how much the correlation matrix is shrunk towards the identity matrix. If $a_P = 1$, the LKJ prior is uniform over correlation matrices of order T, and $a_P > 1$ shrinks the correlation matrix towards the identity.

4.2 Covariance selection with regularized horseshoe priors

Given a $P \times P$ covariance matrix $\mathbf{\Sigma}$ as in (4.4) or (4.9) we propose a selection-shrinkage model for estimating random effects variances and covariances using the modified Cholesky decomposition

$$\mathbf{\Sigma} = \Omega \Gamma \Gamma' \Omega \tag{4.11}$$

where Ω is a diagonal matrix with elements $\omega_p \geq 0$, $p = 1, \ldots, P$, and

$$\Gamma = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \gamma_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{P1} & \gamma_{P2} & \cdots & 1 \end{pmatrix} \tag{4.12}$$

is a lower triangular matrix with 1's on the diagonal and $P(P-1)/2$ free elements γ_{pq}, $p = 2, \ldots, P$, $q = 1, \ldots, p-1$. As a function of the elements of Ω and Γ, the $(p,q)^{th}$ element of $\mathbf{\Sigma}$, σ_{pq}, is

$$\sigma_{pq} = \begin{cases} \omega_p \omega_q \big(\gamma_{pq} + \sum_{k=1}^{q-1} \gamma_{pk}\gamma_{qk}\big) & p > q \\ \omega_p \omega_q \big(\gamma_{qp} + \sum_{k=1}^{p-1} \gamma_{pk}\gamma_{qk}\big) & p < q \\ \omega_p^2 \big(1 + \sum_{k=1}^{p-1} \gamma_{pk}^2\big) & p = q \end{cases} \tag{4.13}$$

for $p, q = 1, \ldots, P$. From (4.13), the correlation $\rho_{pq} = \sigma_{pq}/(\sigma_{pp}\sigma_{qq})$ between random effects p and q is

$$\rho_{pq} = \frac{\gamma_{pq} + \sum_{k=1}^{q-1} \gamma_{pk}\gamma_{qk}}{\sqrt{\Big(1 + \sum_{k=1}^{p-1} \gamma_{pk}^2\Big)\Big(1 + \sum_{k=1}^{q-1} \gamma_{qk}^2\Big)}}. \tag{4.14}$$

for $p > q$. Usually γ_{pq} has the largest impact on the magnitude of ρ_{pq}. This construction of $\mathbf{\Sigma}$ guarantees that $\mathbf{\Sigma}$ is positive-semidefinite when all $\omega_p > 0$.

To model the possibility that certain diagonal elements of Σ might be effectively zero, we propose using the *regularized horseshoe* (RHS) prior (Piironen and Vehtari, 2017) for the elements ω_p of Ω. The RHS prior is a type of *global-local shinkage* prior, where a global shrinkage parameter τ shrinks all elements ω_p towards zero, and local

shrinkage parameters λ_p allow certain ω_p to "escape" the shrinkage. We model ω_p as

$$\begin{aligned}
\omega_p &\sim \mathrm{N}(0, \tau^2\widetilde{\lambda}_p^2)\mathbb{1}_{[\omega_p>0]} \\
\tau &\sim \mathcal{C}^+(0, \tau_0) \\
\widetilde{\lambda}_i &= \frac{c^2\lambda_p}{c^2 + \tau^2\lambda_p^2} \\
\lambda_p &\sim \mathcal{C}^+(0, 1) \\
c^2 &\sim \mathrm{IG}\Big(\frac{v}{2}, \frac{s^2v}{2}\Big),
\end{aligned} \tag{4.15}$$

where τ is the *global shrinkage parameter*, $\widetilde{\lambda}_p$ are *local shrinkage parameters*, $\mathcal{C}^+(0, \tau_0)$ is the half-Cauchy distribution with scale $\tau_0 > 0$ and density

$$\mathcal{C}^+(\tau; 0, \tau_0) \propto (\tau_0^2 + \tau^2)^{-1}\mathbb{1}_{[\tau>0]}, \tag{4.16}$$

c is the *slab width* for elements of Ω, and τ_0, v, and s^2 are known values. The slab width c is the prior standard deviation of ω_p when ω_p is far from zero. The parameter τ is the *global shrinkage parameter* and λ_p are *local shrinkage parameters*. With the RHS prior $\tau^2\lambda_p^2 \ll c^2$ implies the element ω_p is close to zero. When $\tau^2\lambda_p^2 \gg c^2$, the prior (4.15) approaches a half-normal $\mathrm{N}(0, c^2)\mathbb{1}_{[\omega_p>0]}$. The RHS prior can be seen as a continuous alternative to a spike-and-slab prior with finite slab width.

Often we want to shrink correlation elements ρ_{pq} towards zero if either diagonal element σ_{ii} or σ_{jj} is close to zero, for example if there are few studies and there is little information in the data on correlation parameters. The decomposition (4.11), does this by shrinking elements γ_{pq} towards zero if either element ω_p or ω_q is near zero, as ω_p is proportional to σ_{pp} and γ_{pq} generally has the largest impact on the

magnitude of ρ_{pq}. We propose a *conditional shrinkage* (CS) prior $p(\gamma_{pq}|\omega_p, \omega_q)$ for elements γ_{pq}

$$\begin{aligned} \gamma_{pq}|\omega_p, \omega_q &\sim \mathrm{N}(0, \alpha_{pq}^2) \\ \alpha_{pq}^2 &= a_0^2\Big(\frac{1}{\omega_p^2} + \frac{1}{\omega_q^2}\Big)^{-1}, \end{aligned} \tag{4.17}$$

where $a_0 > 0$ is known. In (4.17) the variance α_{pq}^2 of γ_{pq} will be small if either ω_p or ω_q is small and γ_{pq} will be shrunk towards zero. Figure 4.1 is a contour plot of (4.17) as a function of ω_p and ω_q for $a_0 = 4$, showing how the standard deviation α_{pq} of γ_{pq} increases as ω_p and ω_q both deviate from zero. We call the prior for $\mathbf{\Sigma}$ with regularized horseshoe priors (4.15) on ω_p and the conditional shrinkage prior (4.17) on elements γ_{pq} the RHS-CS prior.

If we do not want to shrink correlations ρ_{pq}, we give elements γ_{pq} a normal prior centered at zero with known variance a^2

$$\gamma_{pq} \sim \mathrm{N}(0, a^2). \tag{4.18}$$

We call the model with RHS prior (4.15) on ω_p and prior (4.18) on γ_{pq} (no shrinkage) the RHS-NS prior.

4.2.1 Choosing input values

We need to choose reasonable values for prior parameters τ_0, v, s^2, and a_0^2 in equations (4.15) and (4.17).

The choice of v in (4.15) has little impact on posterior inferences, and we generally choose $v = 2$ so that the prior mean slab width is $\mathrm{E}[c^2|v, s^2] = s^2$. A larger s^2 gives

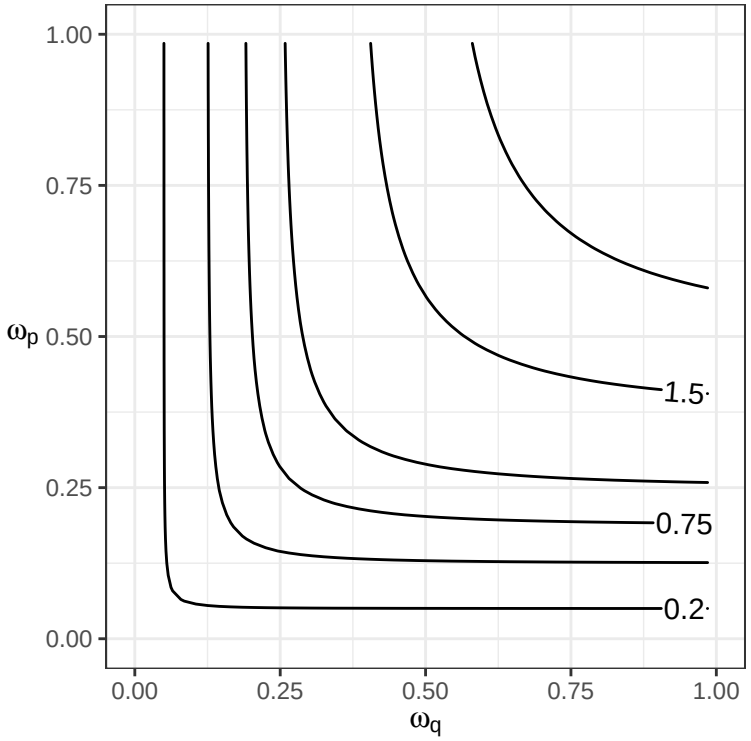

Figure 4.1: Contour plot showing prior standard deviation α_{pq} for γ_{pq} given ω_p, ω_q, and $a_0 = 4$. Contour lines are for $\alpha_{pq} = 0.2$, 0.5, 0.75, 1, 1.5, and 2.

more prior mass to larger values of c^2. We recommend choosing several values of s^2 to see the impact on posterior inferences, and have found that choices near $s^2 = 4$ allow for adequate shrinkage of random effects variances that are near zero while allowing random effects variances far from zero to remain unshrunk. For a_0^2, random effects for both the 3RE and AB-NMA models are on the logit scale and their standard deviations should not be much larger than 1, so parameters ω_p should also not be much larger than 1. Observed correlations become nearly flat on $[-1, 1]$ when one random effects standard deviation is < 0.05, and a prior variance $\alpha_{pq} < 0.2^2$ for γ_{pq} results in $> 80\%$ of the prior mass for $\rho_{pq} \in (-0.2, 0.2)$. This suggests $a_0^2 = 4^2$ as

a reasonable default value, as $\alpha_{pq}^2 = 4^2(1/0.05^2 + 1/1^2)^{-1} \approx 0.2^2$. As with the prior value s^2, we recommend fitting a model with a few different values of a_0. This is because a value of a_0 that is too small will result in shrinkage of correlations even if standard deviations σ_p and σ_q are moderate, while an a_0 that is too large will result in little shrinkage even if σ_p or σ_q is small.

The prior parameter τ_0 is chosen as a measure of how much information there is in the data for the parameters that are modeled with the RHS prior. Piironen and Vehtari (2017) recommend choosing τ_0 based on the expected number of relevant (non-zero) parameters and the sampling variance of the outcome. When the outcome being modeled is normal response data and the parameters being modeled are a set of P regression coefficients, a generic formula for τ_0 is

$$\tau_0 = \frac{p_0}{P - p_0} \frac{\sigma}{N} \tag{4.19}$$

where p_0 is a guess at the expected number of relevant coefficients, σ is the sampling standard deviation of the response variable, and N is the sample size. To our knowledge, RHS priors have not been used for covariance modeling, and we need to modify (4.19) to suit our model. In our case, the outcome data are essentially standard deviations of random effects, and $P = 3$ for the 3RE model and $P = T$ in the AB-NMA model where T is the total number of treatments. To calculate an estimate of how much information there is in the data for the standard deviations of random effects, we replace σ/N with the average standard deviation of the standard deviation of random effects. That is, the average of $\text{SD}(\sigma_\beta)$, $\text{SD}(\sigma_\delta)$, and $\text{SD}(\sigma_\nu)$ in the 3RE model, and the average of $\text{SD}(\sigma_t)$, $t = 1, \ldots, T$ in the AB-NMA model.

To calculate $\text{SD}(\sigma_\beta)$, $\text{SD}(\sigma_\delta)$, and $\text{SD}(\sigma_\nu)$ in the 3RE model, we first calculate

crude estimates $\widehat{\beta}_i$, $\widehat{\delta}_i$, and $\widehat{\nu}_i$ for each study $i = 1, \ldots, S$ as

$$\begin{aligned}
\widehat{\beta}_i &= \frac{1}{2}\left(\log\left(\frac{n_{i11}}{(n_{i1} - n_{i11})}\right) + \log\left(\frac{n_{i01}}{(n_{i0} - n_{i01})}\right)\right) \\
\widehat{\delta}_i &= \log\left(\frac{n_{i11}n_{i00}}{n_{i10}n_{i01}}\right) \\
\widehat{\nu}_i &= \log\left(\frac{n_{i1}}{n_{i0}}\right)
\end{aligned} \tag{4.20}$$

and calculate the sample standard deviations s_β, s_δ, and s_ν of β_i, δ_i, and ν_i across studies. Assuming $\widehat{\beta}_i$, $\widehat{\delta}_i$, and $\widehat{\nu}_i$ are random samples from a normal distribution, we calculate unbiased estimates for the standard deviations $\mathrm{SD}(s_\beta)$, $\mathrm{SD}(s_\delta)$, and $\mathrm{SD}(s_\nu)$ of sample standard deviations s_β, s_δ, and s_ν as

$$\begin{aligned}
\mathrm{SD}(s_\beta) &= s_\beta H(S) \\
\mathrm{SD}(s_\delta) &= s_\delta H(S) \\
\mathrm{SD}(s_\nu) &= s_\nu H(S) \\
H(S) &= \frac{\Gamma(\frac{S-1}{2})}{\Gamma(S/2)}\sqrt{\frac{S-1}{2} - \left(\frac{\Gamma(S/2)}{\Gamma(\frac{S-1}{2})}\right)^2}.
\end{aligned} \tag{4.21}$$

The function $H(S)$ is a correction factor to give an unbiased estimate of $SD(s)$; see Appendix A for details. We then usually set $p_0 = 2$ to indicate that we think one of the random effects has variance near zero, and calculate τ_0 with

$$\tau_0 = \frac{p_0}{3 - p_0}\frac{\mathrm{SD}(s_\beta) + \mathrm{SD}(s_\delta) + \mathrm{SD}(s_\nu)}{3}, \tag{4.22}$$

where the second fraction is the average of $\mathrm{SD}(s_\beta)$, $\mathrm{SD}(s_\delta)$, and $\mathrm{SD}(s_\nu)$.

We follow similar calculations as in (4.20) - (4.22) to calculate τ_0 for the AB-

NMA model. Let $\mathcal{S}_t$ be the set of studies $\mathcal{S} \subseteq \{1, \ldots, S\}$ that report information on treatment t, let $\mathcal{T}_i$ be the set of treatments included in study i, and let $S_t = |\mathcal{S}_t|$ be the number of studies reporting on treatment t and $T_i = |\mathcal{T}_i|$ be the number of treatments in study i. Calculation of τ_0 in the AB-NMA model depends on whether we use the standard linear model (4.7) or the SV linear model (4.10). For the standard model (4.7), for each observed study i and treatment t reported on in study i we calculate $\widehat{\mu}_t + \widehat{\eta}_{it}$ as

$$\widehat{\mu}_t + \widehat{\eta}_{it} = \text{logit}\Big(\frac{y_{it}}{n_{it}}\Big) \tag{4.23}$$

and calculate the observed standard deviation s_t of the $\widehat{\mu}_t + \widehat{\eta}_{it}$ across the S_t studies reporting on treatment t. We then calculate the standard deviation $\text{SD}(s_t)$ of the sample standard deviations for each treatment t as

$$\text{SD}(s_t) = s_t H(S_t) \tag{4.24}$$

where the function $H(\cdot)$ is as defined in (4.21). We then set

$$\tau_0 = \frac{p_0}{T - p_0} \frac{1}{T} \sum_{t=1}^{T} \text{SD}(s_t). \tag{4.25}$$

For the SV model (4.10) we calculate

$$\begin{aligned}
\widehat{\mu}_t &= \frac{1}{S_t} \sum_{i \in \mathcal{S}_t} \text{logit}(y_{it}/n_{it}) \\
\widehat{\eta}_{i0} &= \frac{1}{T_i} \sum_{t \in \mathcal{T}_i} \big(\text{logit}(y_{it}/n_{it}) - \widehat{\mu}_t \big) \\
\widehat{\eta}_{it} &= \text{logit}\big(y_{it}/n_{it}\big) - (\widehat{\mu}_t + \widehat{\eta}_{i0})
\end{aligned} \tag{4.26}$$

and calculate s_t as the standard deviation of $\widehat{\eta}_{it}$ across studies $i \in \mathcal{S}_t$ and calculate $\text{SD}(s_t) = s_t H(S_t)$. We then set τ_0 equal to $p_0/(T - p_0)$ multiplied by the mean of $(\text{SD}(s_1), \ldots, \text{SD}(s_T))$ as in equation (4.25).

We generally prefer to avoid using the data in priors in a Bayesian analysis, and this method described for choosing τ_0 in the RHS prior for the 3RE and AB-NMA models uses the data for setting prior parameters. However, τ_0 as calculated in (4.22) and (4.25) gives only a crude estimate of how precise estimates of random effects standard deviations should be given the number of studies S. In our experiments, perturbing τ_0 by a factor of 0.5 or 2 had a negligible impact on posterior inferences. The local shrinkage parameters $\widetilde{\lambda}_p$ easily outweigh the global shrinkage parameter τ if the data supports random effects standard deviations away from zero.

4.2.2 Incorporating prior information

Sometimes we may have prior information that the correlation between two random effects may be either positive or negative. If we believe the correlation ρ_{pq} between random effects i and j is positive (negative), we can center the prior for γ_{pq} at a known positive (negative) value, such as 0.5 (-0.5).

4.3 Data analyses

We illustrate the 3RE-RHS and AB-NMA-RHS models using both synthetic and real world data examples.

4.3.1 3RE: Synthetic example

We generate a single dataset with $S = 5$ studies. For each study $i = 1, \ldots, S$, the probability π_{i0} of the event given no risk factor is constant at $\pi_{i0} = 0.05$. The probability π_{i1} of the event given the risk factor is varied from $\pi_{i1} = 0.075$ for $i = 1$ and increasing to $\pi_{51} = 0.3$, so that

$$(\pi_{11}, \ldots, \pi_{51})' = (0.075, 0.13125, 0.1875, 0.24375, 0.3)'.$$

The probability ψ_i of the risk factor is held constant at $\psi_i = 0.25$ for all studies i. Sample sizes N_i for each study 1 to 5 are $(N_1, \ldots, N_5)' = (1500, 1000, 2500, 2000, 500)'$, so that sample sizes vary but have little correlation with π_{i1}. Study i's contingency table is drawn from a multinomial distribution with size N_i and with cell probabilities π_{ijk}, $j, k = 0, 1$

$$\begin{aligned}
\pi_{i11} &= \pi_{i1}\psi_i \\
\pi_{i10} &= (1 - \pi_{i1})\psi_i \\
\pi_{i01} &= \pi_{i0}(1 - \psi_i) \\
\pi_{i00} &= (1 - \pi_{i0})(1 - \psi_i).
\end{aligned}$$

The design of probabilities π_{i0} and π_{i1} induces a non-zero variance for both ran-

dom effects β_i and δ_i, true zero variance for $\nu_i = \text{logit}(\psi_i)$. Correlation between β_i and δ_i is strong and positive, and the covariance with the random effect ν_i is zero.

We compare posterior distributions of random effects standard deviations σ_β, σ_δ, and σ_ν and correlations $\rho_{\beta\delta}$, $\rho_{\beta\nu}$, $\rho_{\delta\nu}$ from models with three priors:

1. **IW**: IW(η, I_3) with degrees of freedom $\eta = S + 1$ and an identity input matrix;

2. **LKJ**: A separation strategy with $\Sigma = WPW$, where W is diagonal with elements σ_β, σ_δ, and σ_ν which each have half-Cauchy$(0, 1/\sqrt{2})$ priors, and P is a correlation matrix with an LKJ(1) prior;

3. **RHS-CS**: The Cholesky decomposition $\mathbf{\Sigma} = \Omega\Gamma\Gamma'\Omega$ with RHS prior (4.15) on the diagonal elements ω_k, $k = 1, \ldots, 3$, and the conditional shrinkage prior (4.17) on the lower-triangular free elements of Γ.

For the RHS-CS prior we calculate τ_0 as in Section 4.2.1, and set $v = 4$, $s^2 = 1$, and $a_0 = 3$.

Figure 4.2 shows posterior distributions for random effects standard deviations. Posterior distributions are similar for σ_δ under the three priors. The posterior for σ_β is shifted right slightly for IW compared to LKJ and RHS-CS, and IW has large positive bias for σ_ν while both LKJ and RHS priors yield posteriors with modes near zero. Figure 4.3 shows posterior distributions for the correlations $\rho_{\beta\delta}$ (top), $\rho_{\beta\nu}$ (middle), and $\rho_{\delta\nu}$ (bottom) under the IW (red), LKJ (green), and RHS-CS (blue) priors. The posterior distribution for correlations under the RHS-CS prior looks similar to the IW and LKJ priors for $\rho_{\beta\delta}$ (top) when both random effects β_i and δ_i have variance far from zero. The random effect ν_i has very small variance, and the

RHS-CS prior heavily shrinks $\rho_{\beta\nu}$ (middle) and $\rho_{\delta\nu}$ (bottom) towards zero, while the IW and LKJ priors yield diffuse posterior distributions.

Table 4.1 presents posterior means and credible intervals (CIs), with CIs taken from the 2.5^{th} and 97.5^{th} posterior quantiles, for the *global* contingency table statistics (CTSs) positive and negative likelihood ratios (LR+/-), positive and negative predictive values (PPV/NPV), and sensitivity and specificity. The three priors tend to have similar mean estimates for all CTSs, and the RHS-CS prior yields equal or shorter CI lengths compared to LKJ and strictly shorter CI lengths compared to IW for all CTSs. LR+ has the largest difference in CI lengths between the Models, where IW and LKJ have $\approx 40\%$ and 20% wider CIs than RHS-CS.

CTS	IW	LKJ	RHS-CS
LR−	0.62 (0.43, 0.82)	0.61 (0.43, 0.82)	0.60 (0.44, 0.79)
LR+	2.68 (1.72, 4.45)	2.64 (1.73, 4.00)	2.58 (1.79, 3.71)
NPV	0.94 (0.89, 0.96)	0.94 (0.90, 0.96)	0.94 (0.92, 0.96)
PPV	0.22 (0.13, 0.34)	0.21 (0.13, 0.33)	0.20 (0.14, 0.31)
Sens	0.53 (0.37, 0.67)	0.53 (0.38, 0.66)	0.53 (0.39, 0.66)
Spec	0.77 (0.68, 0.84)	0.78 (0.76, 0.81)	0.78 (0.76, 0.81)

Table 4.1: Posterior summaries of global CTSs for each covariance prior. Each row is a different CTS, and each column represents mean and 95% CI taken as the 2.5^{th} and 97.5^{th} posterior quantiles when modeling the covariance matrix Σ with IW, LKJ, or RHS priors.

4.3.2 3RE: Diagnostic value of risk factors associated with adverse events after syncope

Gibson et al. (2018) fit a 3RE meta-analysis on a set of studies on patients presenting to the emergency department (ED) with syncope. The outcome of interest was 30-day mortality and serious cardiac events, and 32 potential risk factors were analyzed.

Risk Factor	No. Studies	s_β	s_δ	s_ν	τ_0
Chest Pain	3	0.48	0.81	**0.08**	0.42
Male Gender	7	1.05	0.21	**0.05**	0.25
White Race	3	0.74	**0.07**	0.82	0.50

Table 4.2: Values of s_β, s_δ, and s_ν for three risk factors in the syncope data. The final column shows the value of τ_0 used in the RHS prior. Bolded values are < 0.10, signaling that σ_ν may be zero or near-zero for Chest Pain and Male Gender, and σ_δ may be zero or near-zero for White Race.

We look for risk factors that might have zero or near-zero RE variance σ_β^2, σ_δ^2, or σ_ν^2 by calculating s_β, s_δ, and s_ν using equations (4.20) for all risk factors with at least 3 studies. We select risk factors for which at least one of s_β, s_δ, and s_ν was less than 0.10 to illustrate the value of the new RHS-CS prior. The three risk factors male gender, chest pain accompanying syncope, and White race each have one of s_β, s_δ, or $s_\nu < 0.10$, which are shown in Table 4.2 along with the number of studies per risk factor and the value of τ_0 calculated for each risk factor using the method described in Section 4.2.1

We fit models using IW, LKJ, and RHS-CS priors. For each analysis we set 4000 iterations in each of 4 chains, and discard the first 2000 iterations as burn-in. Models are fit in Stan (Gelman et al., 2015) with R (R Core Team, 2021). We compare posterior means and 95% CIs of global CTSs under the three priors.

Table 4.3 gives posterior means and 95% CIs for the CTSs LR−, LR+, NPV, PPV, sensitivity, and specificity for each RF and Model. We see that differences in means are modest, and the RHS-CS Model tends to have a shorter right tail. The RHS-CS prior yields shorter 95% CI lengths than both the IW and LKJ priors for every RF and CTS, with the IW prior generally having the largest CI lengths.

RF	CTS	IW	LKJ	RHS-CS
Chest Pain	LR−	0.97 (0.79, 1.18)	0.96 (0.81, 1.08)	0.97 (0.86, 1.05)
	LR+	1.69 (0.67, 3.92)	1.91 (0.67, 5.78)	1.62 (0.70, 3.80)
	NPV	0.82 (0.62, 0.93)	0.81 (0.56, 0.92)	0.84 (0.65, 0.92)
	PPV	0.19 (0.09, 0.39)	0.21 (0.08, 0.47)	0.19 (0.09, 0.39)
	Sens	0.14 (0.05, 0.32)	0.12 (0.05, 0.28)	0.11 (0.05, 0.22)
	Spec	0.90 (0.77, 0.95)	0.91 (0.84, 0.94)	0.92 (0.88, 0.94)
Male Gender	LR−	0.70 (0.56, 0.85)	0.71 (0.63, 0.78)	0.72 (0.65, 0.78)
	LR+	1.48 (1.22, 1.86)	1.41 (1.31, 1.58)	1.39 (1.31, 1.52)
	NPV	0.93 (0.85, 0.97)	0.92 (0.80, 0.97)	0.93 (0.84, 0.97)
	PPV	0.13 (0.06, 0.27)	0.14 (0.06, 0.29)	0.12 (0.06, 0.25)
	Sens	0.59 (0.48, 0.69)	0.58 (0.54, 0.63)	0.58 (0.55, 0.62)
	Spec	0.58 (0.51, 0.66)	0.59 (0.57, 0.61)	0.58 (0.57, 0.60)
White Race	LR−	0.70 (0.37, 1.38)	0.64 (0.45, 0.99)	0.61 (0.46, 0.82)
	LR+	1.26 (0.94, 1.85)	1.27 (1.04, 1.71)	1.25 (1.08, 1.54)
	NPV	0.93 (0.78, 0.98)	0.92 (0.68, 0.98)	0.94 (0.83, 0.98)
	PPV	0.12 (0.05, 0.32)	0.13 (0.05, 0.40)	0.11 (0.05, 0.26)
	Sens	0.76 (0.52, 0.90)	0.76 (0.50, 0.90)	0.77 (0.56, 0.90)
	Spec	0.34 (0.18, 0.55)	0.36 (0.18, 0.60)	0.35 (0.19, 0.57)

Table 4.3: Results from syncope data analysis. Columns 3-5 give posterior means and 95% CIs for positive and negative likelihood ratios (LR+/-), positive and negative predictive values (PPV/NPV), sensitivity (Sens), and specificity (Spec), for 3 Models using IW, LKJ, and RHS-CS priors for the covariance matrix of random effects.

4.3.3 AB-NMA: Safety of inhaled medications for patients with chronic obstructive pulmonary disease

We re-analyze a dataset of 41 studies on the safety of inhaled medications (SIM) in patients with chronic obstructive pulmonary disease (COPD) first analyzed in Dong et al. (2013). There are 6 treatment arms: tiotropium Soft Mist Inhaler (TIO-SMI), tiotropium HandiHaler (TIO-HH), inhaled corticosteriods (ICS), long-acting $\beta 2$ agonists (LABAs), a LABA-ICS combination, and placebo. Table 4.4 details how many studies reported on each treatment arm. The outcome is all-cause mortality

within a 6-month followup period. There are 31 studies reporting on two treatments, 3 studies reporting on three treatments, and 7 studies reporting on four treatments, with a total of 52462 randomized patients across all treatment arms.

Treatment	No. Observations
TIO-SMI	2
TIO-HH	12
ICS	15
LABA	20
LABA-ICS	17
Placebo	33

Table 4.4: Number of studies reporting data on each of the six treatments in the SIM dataset.

The treatment arm TIO-SMI has only two observations, and in both cases it is in a 2-arm study compared to placebo. While all other treatment arms exhibit heterogeneity in the log-odds of the event across studies, the TIO-SMI arm has near identical event rates of 0.02614 and 0.02617 (log-odds of -3.617 and -3.616), indicating that the random-effects variance for TIO-SMI may be effectively zero. The other five treatment arms have large observed correlations between them; we believe the large correlations may be due to individual study-level random effects, which would suggest fitting the data with linear model (4.10).

We fit five Models to the SIM data:

1. **IW-SM**: IW prior on $\boldsymbol{\Sigma}$ with linear model (4.7),
2. **LKJ-SM**: separation strategy $\boldsymbol{\Sigma} = WPW$ with half-Cauchy priors on diagonal elements of W, LKJ(1) prior on correlation matrix P, and linear model (4.7),
3. **RHS-CS-SM**: RHS-CS prior on $\boldsymbol{\Sigma}$ and linear model (4.7),

4. **RHS-NS-SM**: RHS prior on $\mathbf{\Sigma}$ and linear model (4.7),

5. **RHS-SV**: RHS prior on diagonal $\mathbf{\Sigma}$ and linear model (4.10),

where -SM is short for the Standard Model (4.7) and -SV is the Separate Variance linear model (4.10). We compare Model fit using elpd (Vehtari et al., 2017) with the loo package in R (Vehtari et al., 2020), with inferences for mean treatment effects μ_t, and for absolute risks (ARs) for each treatment arm, where $\text{AR}_t = \text{E}[p_{it}|\mu_t, \sigma_{tt}]$ is the unknown marginal event rate of treatment t and is a function of μ_t and σ_{tt}. Model comparison with elpd is similar to using widely applicable information criterion; elpd $\approx -\frac{1}{2}$WAIC (Gelman et al., 2014) and larger values of elpd indicate better model fit.

Model #	Model	elpd
5	RHS-SV	**-8425.85**
1	IW	-8432.32
4	RHS	-8433.70
3	RHS-CS	-8434.73
2	LKJ	-8435.10

Table 4.5: elpd for each fitted Model in the SIM data analysis. The first column is the Model number; the second column is the Model name, and the third column is elpd. Rows are sorted from largest elpd to smallest. The RHS-SV Model has the largest elpd, indicating better model fit.

Table 4.5 shows elpd for each Model, where we see that linear model (4.10) with RHS prior on $\mathbf{\Sigma}$ has a better fit than all other Models, while the other four SM Models have comparable fit. The elpd difference (95% CI) between RHS-SV the next closest Model (IW) is 6.47 (1.04, 11.90).

Figure 4.4 shows mean treatment effects μ_t for each treatment and Model. All 5 Models have similar mean estimates for each μ_t. The IW and LKJ Models have very large 95% CIs for the treatment TIO-SMI, which only 2 studies reported on,

while the RHS-CS, RHS-NS, and RHS-SV Models have much shorter CI lengths. The RHS-SV Model also has much shorter CI lengths for the treatments TIO-HH and ICS, which have the next fewest studies with 12 and 15, respectively. CI lengths are similar across Models for the other three treatments, which each have at least 17 studies.

Previous implementations of the AB-NMA model (Wang et al., 2021) have used an approximation from Zeger et al. (1988) to calculate AR_t in each iteration of MCMC,

$$\text{AR}_t = \left(1 + \exp\Big(-\mu_t / \sqrt{1 + \frac{256}{76\pi^2}\sigma_{tt}}\Big)\right)^{-1}. \tag{4.27}$$

We instead use a nested Monte Carlo (MC) method to calculate AR_t in each iteration of MCMC. The nested MC method calculates $\text{E}[p_{\text{S}+1,t}|\mu_t, \sigma_{tt}]$, the expected absolute risk for treatment t in a new study $S+1$ given μ_t and σ_{tt}, by approximating the integral

$$\int \frac{1}{1+\exp(-(\mu_t + \eta_{[S+1]t}))} p(\eta_{[S+1],t}|\sigma_{tt}) d\eta_{[S+1]t} \tag{4.28}$$

in each iteration m of MCMC by taking L sub-samples $\eta^{(m,l)}_{[S+1]t}$, $l = 1, \ldots, L$, and calculating a Monte Carlo estimate

$$\text{E}[p_{[S+1]t}|\mu_t^{(m)}, \sigma_{tt}^{(m)}] \approx \frac{1}{L}\sum_{l=1}^{L} \frac{1}{1+\exp(-(\mu_t^{(m)} + \eta^{(m,l)}_{[S+1]t}))} \tag{4.29}$$

where $\mu_t^{(m)}$ and $\sigma_{tt}^{(m)}$ are the m^{th} posterior samples of μ_t and σ_{tt} and $\eta^{(m,l)}_{[S+1]t} \sim \text{N}(0, (\sigma_{tt}^{(m)})^2)$. The nested Monte Carlo method had slightly smaller mean estimates and shifted 95% CIs compared to the approximation (4.27).

Figure 4.5 shows treatment-specific AR posterior means and 95% CIs across Models 1-5 for all treatments. We see that for all treatments except TIO-SMI, posterior mean estimates and 95% CIs are similar for Models 1-4, while Model 5 has shorter CIs for the treatments ICS and TIO-HH. Models 1 and 2 have extremely wide 95% CIs for TIO-SMI, including implausible ARs above 0.1. Models 3-5 all yield wider 95% CIs for TIO-SMI than for the other treatments that had more studies, but the CIs avoid implausibly large ARs.

The SIM dataset illustrates that current priors IW and LKJ for Σ are not informative enough when there are very few studies, and the RHS-CS or RHS-NS priors are promising alternatives. The RHS priors are informative enough to prevent diffuse posterior distributions for ARs when there are few studies, but are not so informative that posteriors are biased for treatment arms with many studies.

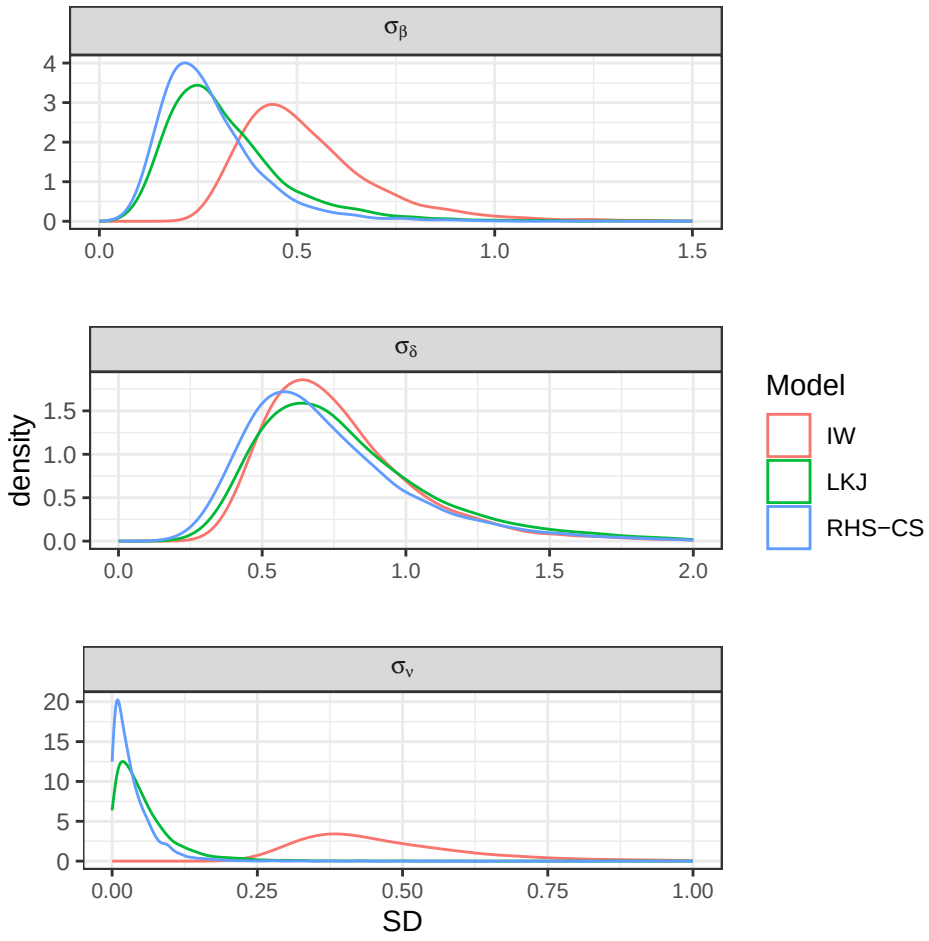

Figure 4.2: Posterior distributions for the SD parameters σ_β (top), σ_δ (middle), and σ_ν (bottom) in the 3RE synthetic data example for three different priors. The red, green, and blue lines are for inverse-Wishart (IW), LKJ, and regularized horseshoe with conditional shrinkage prior (RHS-CS), respectively. All three Models have similar posteriors for σ_δ. The IW Model has positive-shifted posterior distributions for σ_β and σ_ν compared to the LKJ and RHS-CS Models.

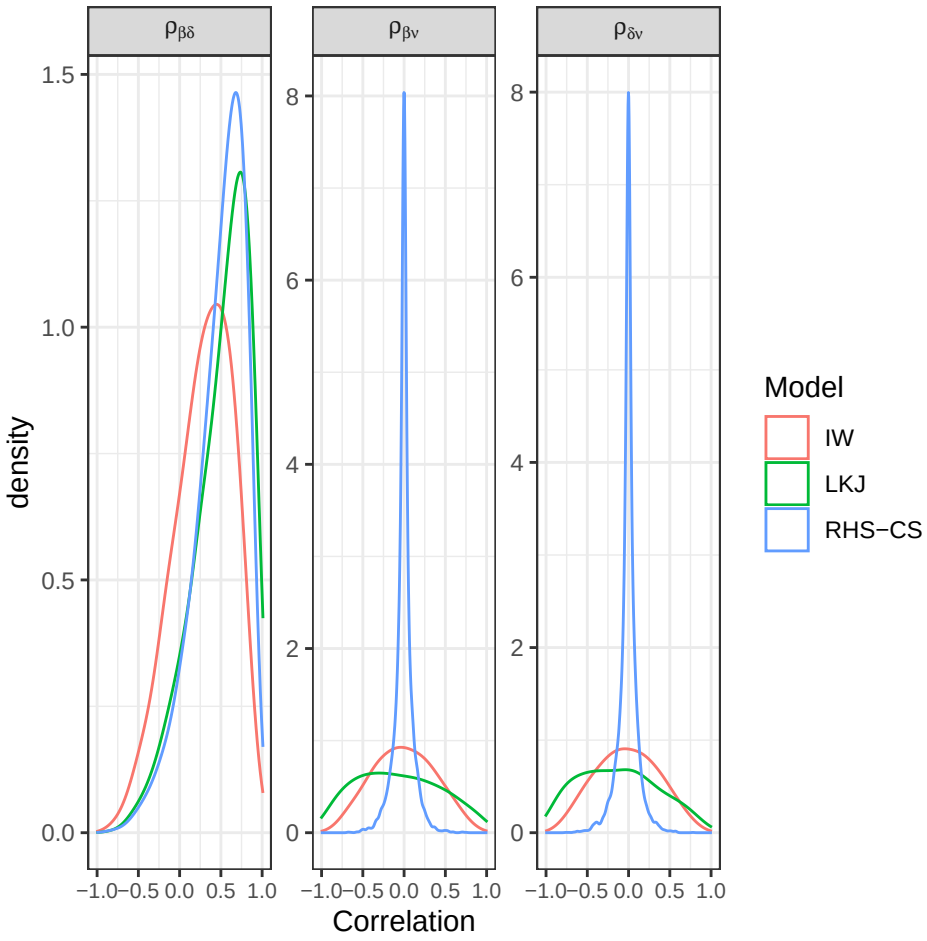

Figure 4.3: Posterior distributions for the correlation parameters $\rho_{\beta\delta}$ (top), $\rho_{\beta\nu}$ (middle), and $\rho_{\delta\nu}$ (bottom) in the 3RE synthetic data example for three different priors. The red, green, and blue lines are for inverse-Wishart (IW), LKJ, and regularized horseshoe with conditional shrinkage prior (RHS-CS), respectively.

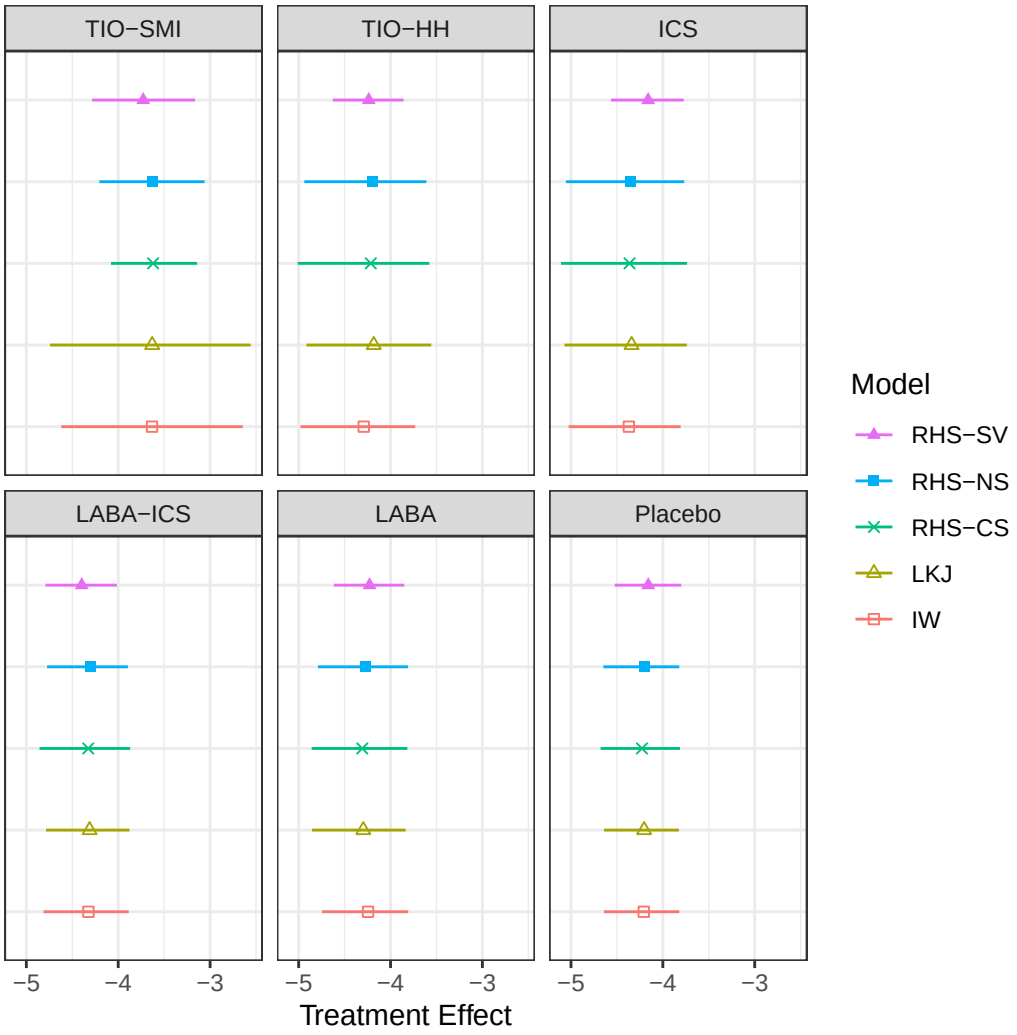

Figure 4.4: Posterior means and 95% CIs for absolute risks (ARs) for the SIM data analysis in Section 4.3.3. Each panel is for a different treatment arm, and Models are differentiated by line color and point shape.

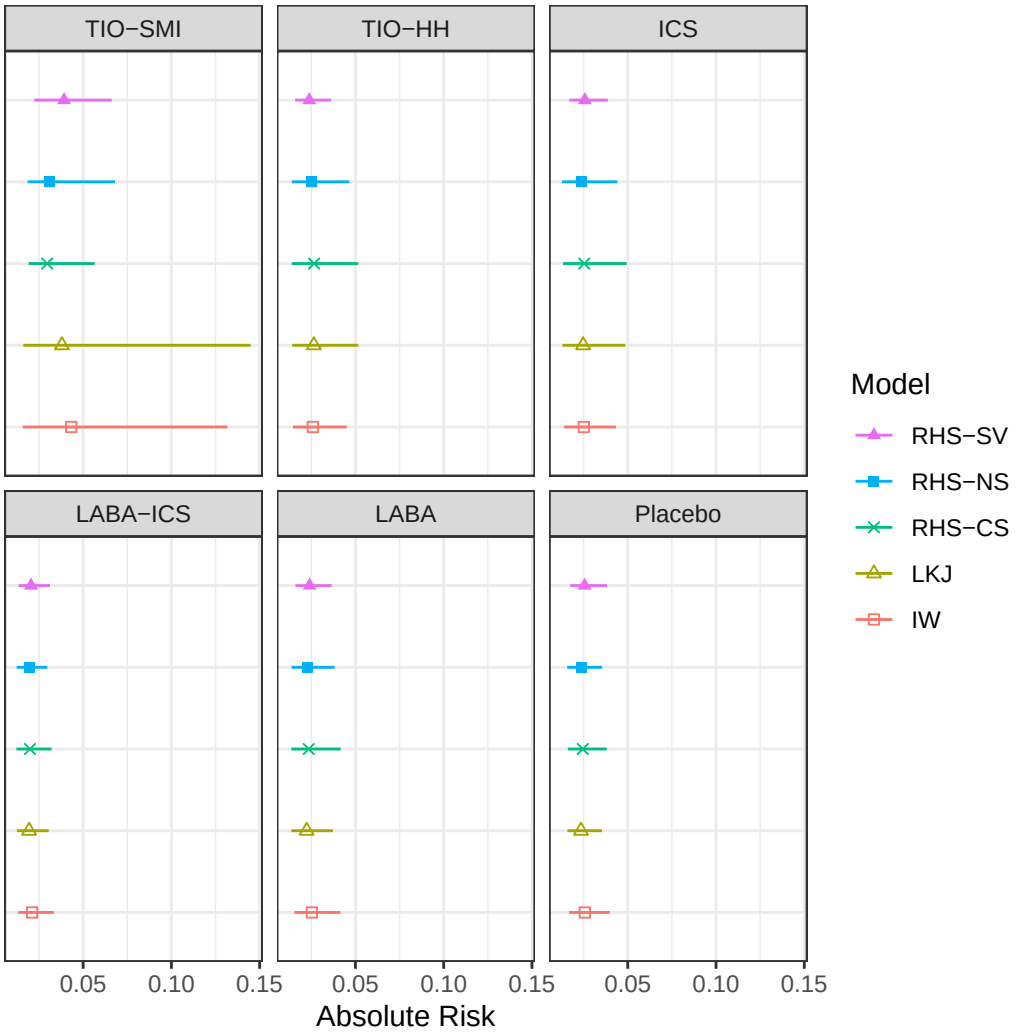

Figure 4.5: Posterior means and 95% CIs for absolute risks (ARs) for the SIM data analysis in Section 4.3.3. Each panel is for a different treatment arm, and Models are differentiated by line color and point shape.

4.4 Discussion

One criticism of using the Cholesky decomposition for modeling covariance matrices has been that marginal priors for variance and correlation parameters depend on their index in the covariance matrix (Wei and Higgins, 2013; Wang et al., 2020). An SD in a higher index has more elements of the Γ matrix contributing to the marginal prior, making the prior less informative. For example, $\sigma_{11} = \omega_1^2$ while $\sigma_{33} = \omega_3^2(1+\gamma_{31}^2+\gamma_{32}^2)$, which is more diffuse in the prior. With more than ~ 5 studies the difference in marginal priors has a negligible impact on posterior distributions. In the AB-NMA data analysis in Section 4.3.3 we ordered treatments by the number of observations in ascending order, giving treatment arms with few studies more informative priors and treatment arms with more studies less informative priors, which is a

We studied 3RE MA and AB-NMA models in this paper, but RHS-based covariance matrix priors could be used in other meta-analysis models, including CB-NMA models (Lu and Ades, 2004; Dias et al., 2013) and Copas models for publication bias (Mavridis et al., 2013, 2014). The CB-NMA model assumes that log-odds ratios δ_{it}, $t = 2, \ldots, T$, relative to a baseline treatment ($t = 1$) are exchangeable, and models

$$\begin{aligned} \text{logit}(p_{it}) &= \mu_i + \delta_{it} \\ \boldsymbol{\delta}_i &\sim \text{N}(\mathbf{0}_{T-1}, \boldsymbol{\Sigma}) \end{aligned}$$

where μ_i is the probability of the event in the baseline treatment group in study i, $\boldsymbol{\delta_i} = (\delta_{2i}, \ldots, \delta_{Ti})'$, and $\boldsymbol{\Sigma}$ is the covariance matrix for random effects δ_{it} containing the variances and correlations of log-odds ratios for each treatment relative to the reference treatment. If we believe that a certain treatment effect δ_{it} should be

constant across studies, or that certain treatment effects should be uncorrelated, the RHS-NS or RHS-CS prior would be appropriate choices.

Selection models for publication bias measure the correlation between observed effects and a latent variable which is the "propensity for publication", with zero correlation implying no publication bias and large correlation implying severe publication bias. In a study design d comparing T_d treatments, there are $\binom{T_d}{2}$ correlation parameters to measure. For example, say design $d = 1$ compares $T_1 = 3$ treatments, labeled A, B, and C, and S_1 studies indexed by $i = 1, \ldots, S_1$ studies have design $d = 1$. Treatment effects are contrasts y_{i1}^{AB} and y_{i1}^{AC} with associated standard errors s_{i1}^{AB} and s_{i1}^{AC}, and measured covariance $c_{id} = \text{cov}(y_{id}^{AB}, y_{id}^{AC})$ reported by studies $i = 1, \ldots, S_1$. A latent variable z_{i1} with marginal mean u_i representing the propensity for publication is modeled with the contrasts as

$$\begin{pmatrix} y_{i1}^{AB} \\ y_{i1}^{AC} \\ z_{i1} \end{pmatrix} \mid \begin{pmatrix} \rho_1^{AB} \\ \rho_1^{AC} \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} \theta_{i1}^{AB} \\ \theta_{i1}^{AC} \\ u_{id} \end{pmatrix}, \begin{pmatrix} (s_{i1}^{AB})^2 & c_{id} & \rho_1^{AB} s_{i1}^{AB} \\ c_{id} & (s_{i1}^{AC})^2 & \rho_1^{AC} s_{i1}^{AC} \\ \rho_1^{AB} s_{i1}^{AB} & \rho_1^{AC} s_{i1}^{AC} \sigma_\nu & 1 \end{pmatrix} \right) \mathbb{1}_{z_{i1}>0} \quad (4.30)$$

where large ρ_1^{AB} or ρ_1^{AC} implies that the probability of publication is strongly related to observed effects y_{i1}^{AB} or y_{i1}^{AC} relative to their means θ_{i1}^{AB} and θ_{i1}^{AC}. To model the belief that publication bias may not be present for all designs and treatment comparisons, a RHS prior could be used to induce sparsity in the covariance matrices for each design d by transforming correlations $\rho_d^{(\cdot,\cdot)}$ with Fisher's z-transformation and placing RHS priors directly on the transformed correlations.

APPENDIX A

Standard deviation of sample standard deviation

Let $x_1, \ldots, x_n$ be a sample of size n from a normal distribution $\mathrm{N}(\mu, \sigma^2)$. Define the sample variance s^2 and sample standard deviation s in the usual way as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$
$$s = \sqrt{s^2}$$

where $\overline{x}$ is the sample mean. We want the standard deviation of s,

$$\begin{aligned} \mathrm{SD}(s) &= \sqrt{\mathrm{Var}(s)} \\ &= \sqrt{\mathrm{E}[s^2] - \mathrm{E}[s]^2}. \end{aligned} \tag{A.1}$$

Because s^2 is unbiased, we have that $\mathrm{E}[s^2] = \sigma^2$. From Gurland and Tripathi (1971) we have that

$$\mathrm{E}[s] = \sqrt{\frac{2\sigma^2}{n-1}}\Big(\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}\Big) \tag{A.2}$$

$$\Rightarrow \quad \mathrm{SD}(s) = \sqrt{\mathrm{E}[s^2] - \mathrm{E}[s]^2} \tag{A.3}$$

$$= \sigma\sqrt{1 - \frac{2}{n-1}\Big(\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}\Big)^2}. \tag{A.4}$$

To calculate an unbiased estimate of SD(s) we plug in an unbiased estimate

$$\hat{\sigma} = s\Big(\frac{n-1}{2}\Big)^{1/2}\frac{\Gamma((n-1)/2)}{\Gamma(n/2)}. \tag{A.5}$$

for σ in (A.4) to get

$$\begin{aligned}\mathrm{SD}(s) &= s\left[\frac{\Gamma((n-1)/2)}{\Gamma(n/2)}\sqrt{\frac{n-1}{2} - \Big(\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}\Big)^2}\right] \\ &= sH(n),\end{aligned} \tag{A.6}$$

which is the formula we use in Section 4.2.1 to calculate the standard deviation of random effect standard deviations.

Bibliography

Arends, L., Hamza, T., Van Houwelingen, J., Heijenbrok-Kal, M., Hunink, M., and Stijnen, T. (2008). Bivariate random effects meta-analysis of ROC curves. *Medical Decision Making* **28,** 621–638.

Bai, R., Lin, L., Boland, M. R., and Chen, Y. (2020). A robust Bayesian Copas selection model for quantifying and correcting publication bias. *arXiv preprint arXiv:2005.02930* .

Banerjee, S. and Ghosal, S. (2015). Bayesian structure learning in graphical models. *Journal of Multivariate Analysis* **136,** 147–162.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10,** 1281–1311.

Begg, C. B. and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics* **50,** 1088–1101.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian Theory*. Wiley & Sons.

Birnbaum, A., Esses, D., Bijur, P., Wollowitz, A., and Gallagher, E. J. (2008). Failure to validate the San Francisco Syncope rule in an independent emergency department population. *Annals of Emergency Medicine* **52,** 151–159.

Bornmann, L., Mutz, R., and Daniel, H.-D. (2007). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics* **1,** 226–238.

Cai, B. and Dunson, D. B. (2006). Bayesian covariance selection in generalized linear mixed models. *Biometrics* **62,** 446–457.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59,** 762–769.

Chu, H. and Cole, S. R. (2006). Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology* **59,** 1331.

Chu, H., Nie, L., Chen, Y., Huang, Y., and Sun, W. (2012). Bivariate random effects models for meta-analysis of comparative studies with binary outcomes: methods for the absolute risk difference and relative risk. *Statistical Methods in Medical Research* **21,** 621–633.

Chu, H., Nie, L., Cole, S. R., and Poole, C. (2009). Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parameterizations and model selection. *Statistics in Medicine* **28,** 2384–2399.

Clyde, M. and Iversen, E. S. (2013). Bayesian model averaging in the M-open framework. In *Bayesian Theory and Applications*, chapter 24, pages 483–498. Oxford University Press.

Copas, J. (1999). What works?: selectivity models and meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **162,** 95–109.

Copas, J. and Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics* **1,** 247–262.

Copas, J. and Shi, J. Q. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research* **10,** 251–265.

Cripps, E., Carter, C., and Kohn, R. (2005). Variable selection and covariance selection in multivariate regression models. *Handbook of Statistics* **25,** 519–552.

Deeks, J. J. and Altman, D. G. (2004). Diagnostic tests 4: likelihood ratios. *British Medical Journal* **329,** 168–169.

Dempster, A. P. (1972). Covariance selection. *Biometrics* **28,** 157–175.

Dias, S., Sutton, A. J., Ades, A., and Welton, N. J. (2013). Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making* **33,** 607–617.

Dong, Y.-H., Lin, H.-H., Shau, W.-Y., Wu, Y.-C., Chang, C.-H., and Lai, M.-S. (2013). Comparative safety of inhaled medications in patients with chronic obstructive pulmonary disease: systematic review and mixed treatment comparison meta-analysis of randomised controlled trials. *Thorax* **68,** 48–56.

Duval, S. and Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* **56,** 455–463.

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* **315,** 629–634.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9,** 432–441.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1,** 515–534.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing* **24,** 997–1016.

Gelman, A., Lee, D., and Guo, J. (2015). Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics* **40,** 530–543.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88,** 881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7,** 339–373.

Gibson, T. A., Weiss, R. E., and Sun, B. C. (2018). Predictors of short-term outcomes after syncope: a systematic review and meta-analysis. *Western Journal of Emergency Medicine* **19,** 517.

Givens, G. H., Smith, D., and Tweedie, R. (1997). Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science* **12,** 221–250.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102,** 359–378.

Guan, M. and Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin & Review* **23,** 74–86.

Guo, J., Riebler, A., and Rue, H. (2017). Bayesian bivariate meta-analysis of diagnostic test studies with interpretable priors. *Statistics in Medicine* **36,** 3039–3058.

Gurland, J. and Tripathi, R. C. (1971). A simple approximation for unbiased estimation of the standard deviation. *The American Statistician* **25,** 30–32.

Hackshaw, A. K., Law, M. R., and Wald, N. J. (1997). The accumulated evidence on lung cancer and environmental tobacco smoke. *British Medical Journal* **315,** 980–988.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics* **9,** 61–85.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science* **7,** 246–255.

Higgins, J. P., Thompson, S. G., and Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172,** 137–159.

Hong, H., Chu, H., Zhang, J., and Carlin, B. P. (2016). A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research Synthesis Methods* **7,** 6–22.

Hoyer, A. and Kuss, O. (2018). Meta-analysis for the comparison of two diagnostic tests to a common gold standard: a generalized linear mixed model approach. *Statistical Methods in Medical Research* **27,** 1410–1421.

Ishwaran, H., Rao, J. S., et al. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics* **33,** 730–773.

Iyengar, S. and Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science* **3,** 109–117.

Jin, Z.-C., Zhou, X.-H., and He, J. (2015). Statistical methods for dealing with publication bias in meta-analysis. *Statistics in Medicine* **34,** 343–360.

Kicinski, M., Springate, D. A., and Kontopantelis, E. (2015). Publication bias in meta-analyses from the Cochrane Database of Systematic Reviews. *Statistics in Medicine* **34,** 2781–2793.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B* **60,** 65–81.

Landenberger, N. A. and Lipsey, M. W. (2005). The positive effects of cognitive–behavioral programs for offenders: A meta-analysis of factors associated with effective treatment. *Journal of Experimental Criminology* **1,** 451–476.

Le, T. and Clarke, B. (2017). A Bayes interpretation of stacking for M-complete and M-open settings. *Bayesian Analysis* **12,** 807–829.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* **100,** 1989–2001.

Lian, Q., Hodges, J. S., and Chu, H. (2019). A Bayesian Hierarchical Summary Receiver Operating Characteristic Model for network meta-analysis of diagnostic tests. *Journal of the American Statistical Association* **114,** 949–961.

Light, R. J. and Pillemer, D. B. (1984). *Summing Up: The Science of Reviewing Research.* Harvard University Press.

Lin, L. and Chu, H. (2018). Quantifying publication bias in meta-analysis. *Biometrics* **74,** 785–794.

Lu, G. and Ades, A. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* **23,** 3105–3124.

Ma, X., Lian, Q., Chu, H., Ibrahim, J. G., and Chen, Y. (2018). A Bayesian hierarchical model for network meta-analysis of multiple diagnostic tests. *Biostatistics* **19,** 87–102.

Ma, X., Nie, L., Cole, S. R., and Chu, H. (2016). Statistical methods for multivariate meta-analysis of diagnostic tests: an overview and tutorial. *Statistical Methods in Medical Research* **25,** 1596–1619.

Macaskill, P., Walter, S. D., and Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine* **20,** 641–654.

Maier, M., Bartoš, F., and Wagenmakers, E.-J. (2022). Robust Bayesian meta-analysis: Addressing publication bias with model-averaging. *Psychological Methods, Advance online publication. https://doi.org/10.1037/met0000405* .

Mavridis, D., Sutton, A., Cipriani, A., and Salanti, G. (2013). A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. *Statistics in Medicine* **32,** 51–66.

Mavridis, D., Welton, N. J., Sutton, A., and Salanti, G. (2014). A selection model

for accounting for publication bias in a full network meta-analysis. *Statistics in Medicine* **33,** 5399–5412.

Moses, L. E., Shapiro, D., and Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in Medicine* **12,** 1293–1316.

Ning, J., Chen, Y., and Piao, J. (2017). Maximum likelihood estimation and EM algorithm of Copas-like selection model for publication bias correction. *Biostatistics* **18,** 495–504.

Owen, D. B. (1980). A table of normal integrals. *Communications in Statistics-Simulation and Computation* **9,** 389–419.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., and Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *Journal of the American Medical Association* **295,** 676–680.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics* **11,** 5018–5051.

Plummer, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria.

Probst, M. A., Gibson, T., Weiss, R. E., Yagapen, A. N., Malveau, S. E., Adler, D. H., Bastani, A., Baugh, C. W., Caterino, J. M., Clark, C. L., Diercks, D. B., Hollander, J. E., Nicks, B. A., Nishijima, D. K., Shah, M. N., Stiffler, K. A.,

Storrow, A. B., Wilber, S. T., and Sun, B. C. (2020). Risk stratification of older adults who present to the emergency department with syncope: the FAINT score. *Annals of Emergency Medicine* **75,** 147–158.

Probst, M. A., Kanzaria, H. K., Gbedemah, M., Richardson, L. D., and Sun, B. C. (2015). National trends in resource utilization associated with ED visits for syncope. *The American Journal of Emergency Medicine* **33,** 998–1001.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ranganathan, P. and Aggarwal, R. (2018). Understanding the properties of diagnostic tests–part 2: Likelihood ratios. *Perspectives in Clinical Research* **9,** 99–102.

Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M., and Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* **58,** 982–990.

Rothstein, H. R., Sutton, A. J., and Borenstein, M. (2006). *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments*. Wiley & Sons.

Rutter, C. M. and Gatsonis, C. A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* **20,** 2865–2884.

Shi, L. and Lin, L. (2019). The trim-and-fill method for publication bias: practical guidelines and recommendations based on a large database of meta-analyses. *Medicine* **98,** https://doi.org/10.1097/MD.0000000000015987.

Smith, T. C., Spiegelhalter, D. J., and Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* **14,** 2685–2699.

Sterne, J. A., Becker, B. J., Egger, M., et al. (2005). The funnel plot. In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, chapter 5, pages 73–98. Wiley & Sons.

Takagi, H., Sekino, S., Kato, T., Matsuno, Y., and Umemoto, T. (2006). Revisiting evidence on lung cancer and passive smoking: adjustment for publication bias by means of "trim and fill" algorithm. *Lung Cancer* **51,** 245–246.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2020). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.4.1.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27,** 1413–1432.

Vevea, J. L. and Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika* **60,** 419–435.

Vevea, J. L. and Woods, C. M. (2005). Publication bias in research synthesis: sensitivity analysis using a priori weight functions. *Psychological Methods* **10,** 428–443.

Wang, Z., Lin, L., Hodges, J. S., and Chu, H. (2020). The impact of covariance priors on arm-based Bayesian network meta-analyses with binary outcomes. *Statistics in Medicine* **39,** 2883–2900.

Wang, Z., Lin, L., Hodges, J. S., MacLehose, R., and Chu, H. (2021). A variance shrinkage method improves arm-based Bayesian network meta-analysis. *Statistical Methods in Medical Research* **30,** 151–165.

Wei, Y. and Higgins, J. P. (2013). Bayesian multivariate meta-analysis with multiple outcomes. *Statistics in Medicine* **32,** 2911–2934.

Wong, F., Carter, C. K., and Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika* **90,** 809–830.

Wynants, L., Riley, R., Timmerman, D., and Van Calster, B. (2018). Random-effects meta-analysis of the clinical utility of tests and prediction models. *Statistics in Medicine* **37,** 2034–2052.

Yao, Y., Pirš, G., Vehtari, A., and Gelman, A. (2021). Bayesian hierarchical stacking. *arXiv preprint arXiv:2101.08954* .

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis* **13,** 917–1007.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44,** 1049–1060.

Zhang, J., Carlin, B. P., Neaton, J. D., Soon, G. G., Nie, L., Kane, R., Virnig, B. A., and Chu, H. (2014). Network meta-analysis of randomized clinical trials: reporting the proper summaries. *Clinical Trials* **11,** 246–262.

Zhang, J., Chu, H., Hong, H., Virnig, B. A., and Carlin, B. P. (2017). Bayesian hierarchical models for network meta-analysis incorporating nonignorable missingness. *Statistical Methods in Medical Research* **26,** 2227–2243.