BAYESIAN HIERARCHICAL MODELS FOR META-ANALYSIS

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

LIANNE K. SIEGEL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

ADVISED BY DR. HAITAO CHU

April, 2021

© LIANNE K. SIEGEL 2021
ALL RIGHTS RESERVED

Acknowledgements

This dissertation would not have been possible without the support of my friends and family. I would especially like to thank my advisor, Dr. Haitao Chu, who has helped me develop into an independent biostatistician. I am very grateful to have had such a supportive mentor over the past four years; thank you for encouraging me to think critically and pushing me to grow as a researcher by presenting at conferences, reviewing journal articles, and starting to co-mentor other students. I look forward to continuing our always longer than anticipated conversations as colleagues. I would also like to thank the University of Minnesota Division of Biostatistics for supporting me with a fellowship during my first two years, allowing me to begin working with Dr. Chu. Thank you as well to Dr. James Neaton and the National Heart, Lung, and Blood Institute (T32HL129956) for supporting my work during my third and fourth years and providing invaluable training in conducting clinical trials. I would also like to thank Dr. M. Hassan Murad, Dr. Richard Riley, Dr. Zhen Wang, and Dr. Fateh Bazerbachi for their contributions to the work estimating reference ranges, as well as Dr. Kyle Rudser and the PLUS consortium for their contributions to Chapter 2 of this dissertation. Finally, I would like to thank Dr. Thomas Murray, Dr. Eric Lock, and Dr. Richard MacLehose for serving on my committee and providing valuable feedback on my work, and Dr. James Hodges for his feedback on Chapter 3.

Dedication

This dissertation is dedicated to my parents, Scot and Debra Siegel, whose endless support of my education has made this possible.

Abstract

Meta-analysis is an important and widely used tool for synthesizing information from multiple independent but related studies. While many meta-analyses, such as those of randomized controlled trials, focus on the synthesis of treatment effects across studies, this dissertation will focus on the meta-analysis of prevalence and normative data. The first part of this thesis concerns the multivariate meta-analysis of prevalence data. When conducting a meta-analysis involving prevalence data for an outcome with several subtypes, each of them is typically analyzed separately using a univariate meta-analysis model. Recently, multivariate meta-analysis models have been shown to correspond to a decrease in bias and variance for multiple correlated outcomes compared with univariate meta-analysis, when some studies only report a subset of the outcomes. Chapter 2 of this thesis proposes a novel Bayesian multivariate random effects model to account for the natural constraint that the prevalence of any given subtype cannot be larger than that of the overall prevalence. Extensive simulation studies show that this new model can reduce bias and variance when estimating subtype prevalences in the presence of missing data, compared with standard univariate and multivariate random effects models. The data from a rapid review on occupation and lower urinary tract symptoms by the Prevention of Lower Urinary Tract Symptoms Research Consortium are analyzed as a case study to estimate the prevalence of urinary incontinence and several incontinence subtypes among women in suspected high risk work environments.

The second part of this thesis concerns estimating a reference range from a meta-analysis. Clinicians frequently must decide whether a patient's measurement reflects that of a healthy "normal" individual. Thus, the reference range is defined as the interval in which some proportion (frequently 95%) of measurements from a healthy population is expected to fall. One can estimate it from a single study, or preferably from a meta-analysis of multiple

studies to increase generalizability. This range differs from the confidence interval for the pooled mean or the prediction interval for a new study mean in a meta-analysis, which do not capture natural variation across healthy individuals. Chapter 3 proposes three methods for estimating the reference range from a meta-analysis of aggregate data that incorporate both within and between-study variations. The results of a simulation study are presented demonstrating that the methods perform well under a variety of scenarios, though users should be cautious when the number of studies is small and between-study heterogeneity is large. These methods are applied to two examples: pediatric time spent awake after sleep onset and frontal subjective postural vertical measurements. Chapter 4 provides a guide for clinicians and epidemiologists explaining the three approaches for estimating the reference range presented in Chapter 3: a frequentist, a Bayesian, and an empirical method. Each method is also extended to individual participant data (IPD) meta-analysis, with the latter being the gold standard when available. These approaches are illustrated using a clinical scenario about the normal range of a liver stiffness test.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Meta-analysis allows for the synthesis of results across multiple independent studies and therefore plays an important role in evidence-based medicine [1, 2]. The number of meta-analyses published has increased sharply in the past several decades, increasing by 20-fold between 1994 and 2014 [2]. New methods developed for meta-analysis include those for network meta-analysis, which allows for the synthesis of studies assessing multiple treatments, and multivariate meta-analysis, which allows for multiple outcomes [3]. In addition, many methods have been developed for the evaluation of diagnostic tests [4–6]. However, while meta-analyses often include the results of clinical trials or other experimental data, there are many examples of meta-analyses of observational data, including prevalence and normative data [7–14]. Normative data refers to data assumed to be drawn from a predefined healthy population that can provide a reference when determining whether the measurements or results in a new population are normal or not [15]. This thesis focuses on methods for meta-analysis of prevalence data and normative data.

The second chapter of this thesis demonstrates how an arm-based network meta-analysis model can be applied to the multivariate meta-analysis of prevalence data. It also proposes a new parameterization that accounts for natural constraints in the data, decreasing bias and increasing precision. There has been much work investigating the benefits of the joint

meta-analysis of multiple related outcomes, particularly in the presence of missing data, due to the multivariate models "borrowing strength" across outcomes [3, 16]. However, there have only been several previous examples of multivariate meta-analysis of prevalence data in the literature [10,11,17], none of which have accounted for natural constraints in the underlying prevalences and observed counts when several outcomes are subsets of an overall outcome. While Trikalinos et al. [18] proposed a multivariate meta-analysis model using a multinomial distribution that could be applied for multiple category prevalences if the outcomes were mutually exclusive, many prevalence outcomes with multiple categories are not mutually exclusive, such as the proportion of individuals in a study with different types of food allergies [13]. This is also true for the motivating example described in Chapter 2, which consists of the results of a systematic review investigating the prevalence of women in high risk working environments with different non-mutually exclusive types of urinary incontinence: stress urinary incontinence and urgency urinary incontinence [19].

Chapter 3 proposes methods for establishing reference ranges for continuous measurements from meta-analyses of normative data and illustrates them through several case studies. A reference range is defined as an interval that captures some predefined proportion of measurements (such as 95%) from a healthy population that can serve as a reference for future comparison. Alternatively, it can be defined as a prediction interval for the measurement of a new individual [20, 21]. Previously, no guidance has existed in the literature for how to estimate a reference range from a meta-analysis, particularly when only aggregate data are available for each study. When conducting a meta-analysis to estimate a reference range, practitioners can expect to have information on the observed mean, standard deviation, and sample size of measurements from each study. There may also be some aggregate demographic information, such as the proportion of males and females in the study, or the mean age of participants. However, individual participant data (IPD) often are not available. Therefore, it is important to develop methods that allow for estimating a reference range based only on the aggregate data. Chapter 3 will propose three such methods: one

frequentist, one Bayesian, and one empirical, and demonstrate their performance through simulations.

Finally, while Chapter 3 proposes three methods to estimate the reference range from a meta-analysis, filling an important gap in the literature, it is aimed at a statistical, rather than clinical audience. Therefore, in Chapter 4, we provide a guide for clinicians and epidemiologists that introduces practitioners to the three methods proposed in Chapter 3. This also demonstrates how these methods can be extended in the case IPD are available. The differences between a reference range, a confidence interval for the pooled mean, and the prediction interval for a new study mean are further explained, as some practitioners have previously reported the pooled mean and its corresponding confidence interval as a "reference value." This chapter also considers ideas such as heterogeneity and applicability. When estimating a reference range, it is important to carefully consider the target population and establish inclusion and exclusion criteria that ensure the reference range will apply to the population of interest. However, if the population of interest consists of several distinct subgroups with different normal measurements, separate reference ranges for these subgroups would likely be more informative. These concepts are illustrated through a clinical scenario regarding the normal range for a non-invasive liver stiffness test.

Chapter 5 summarizes the findings in the previous chapter and describes future work related to these topics.

Chapter 2

A Bayesian Multivariate Meta-Analysis of Prevalence Data

2.1 Introduction

Meta-analysis plays an important role in synthesizing evidence from multiple sources, supporting the recent rapid growth of "evidence-based medicine" [1]. Multivariate and network meta-analysis (NMA) methods have been developed for meta-analyses of data consisting of multiple outcomes, multiple treatments, or multiple diagnostic tests [3, 6, 18, 22–30]. NMA uses both indirect and direct comparisons of multiple treatments within a network, while multivariate meta-analysis allows for the joint analysis of multiple outcomes by incorporating information about their correlations [25, 27]. These models are therefore able to use information normally unavailable when each treatment or outcome is analyzed separately, a statistical concept known as "borrowing strength" that is particularly useful in the presence of missing data [16, 24, 31]. For example, Williams and Bürkner (2017) [32] jointly modeled the effects of intranasal oxytocin on multiple symptoms of schizophrenia in a Bayesian multivariate meta-analysis, resulting in increased precision compared to previous analyses that modeled symptoms separately.

Multivariate meta-analysis models have also been increasingly used in the evaluation of diagnostic tests, as they allow for the joint modeling of accuracy indices such as sensitivity and specificity [4, 5, 33–35]. The use of NMA in meta-analyses of clinical trials assessing multiple treatments has also increased sharply over the past decade [27, 36, 37]. However, many practitioners conducting systematic reviews involving multiple correlated outcomes of interest still analyze these individually using separate univariate models, thus ignoring any correlations between outcomes. Riley (2009) [16] suggested that reasons for this hesitancy may include "tradition, the increased complexity of the multivariate approach, the need for speciali[zed] statistical software and a lack of understanding of the consequences of ignoring correlation in meta-analysis", which are all still relevant issues.

This widespread use of separate univariate models is especially prevalent in the context of observational data. For instance, we found only three examples of multivariate meta-analysis of observational data in the literature: 1) Fawcett et al. (2018) [10] presented a multivariate method that used data on prevalences of individual disorders in order to estimate their overall prevalence. To our knowledge, this is the only study to perform a multivariate meta-analysis of prevalence and incidence data. 2) The Fibrinogen Studies Collaboration (2009) [11] used a bivariate random effects meta-analysis in order to jointly model partially and fully adjusted estimates of the association between fibrinogen level and incidence of coronary heart disease, and 3) Lin and Chu (2018) [17] proposed a Bayesian multivariate meta-analysis simultaneously analyzing multiple factors. While meta-analyses are most frequently conducted in order to estimate effect sizes such as odds ratios (OR's), risk differences, or mean differences, they can also be used to estimate the pooled disease frequency such as incidence rates and prevalence proportions [38]. This may include multiple related outcomes consisting of an overall prevalence and several subtypes of the measured outcome. For example, Rona et al. (2007) [13] conducted univariate meta-analyses estimating the prevalence of any food allergy as well as the prevalences of specific types such as allergies to peanuts and shellfish. Similarly, Williams et al. (2006) [14] separately estimated

the prevalences of typical autism and all autism spectrum disorders (ASD). For this type of data, it is reasonable to expect that the prevalence of any given subtype will be both correlated with and constrained by the overall prevalence. Therefore, these outcomes are not independent. Additionally, the subtypes measured in observational data may be particularly susceptible to outcome reporting bias (ORB). If a study is smaller or performed in a population in which the overall prevalence is expected to be lower, investigators may be more likely to only report the overall outcome [18].

Multivariate meta-analysis models have been shown to be effective in reducing ORB when there is correlation between outcomes, in addition to providing more precise estimates in the presence of missing data [16, 18, 27, 31, 39]. Modeling multiple correlated prevalences using univariate models ignores studies in which each particular prevalence is not reported. This can result in biased estimates if any of the studies are subject to ORB. Instead, multivariate models allow us to "borrow" information on missing observations across outcomes using the within-study correlations [27]. However, one reason that practitioners may be reluctant to use multivariate meta-analysis methods, is that these within-study correlations have to be estimated [3]. In this paper, we address that issue through the use of a Bayesian multivariate meta-analysis framework. This framework also gives us greater flexibility in parameterizing the multivariate random effects model. Prevalences of individual subtypes are subject to the natural constraint that they cannot be larger than the overall prevalence. While Trikalinos et al. (2013) [18] jointly modeled multiple categorical outcomes that are mutually exclusive or subsets of each other using a multinomial distribution, our method differs in that the multiple subtypes modeled need not imply a set of mutually exclusive categories. We introduce a case study as a motivating example in Section 2.2. We then present fully Bayesian univariate and multivariate models for estimating the prevalence of each outcome in Section 2.3, including a novel parameterization of the multivariate random effects model that accounts for the natural constraints in the data, thereby incorporating additional information into the model. We then compare these three different approaches in

simulation studies (Section 2.5) and when applied to the case study (Section 2.6). Section 2.6.2 presents a sensitivity analysis of the missing at random (MAR) assumption for the case study. Section 2.7 gives a brief discussion of these results.

2.2 A Motivating Study

Recently, members of the PLUS research consortium [40] conducted a rapid review of studies reporting lower urinary tract symptoms in women in suspected high risk working environments [19]. Of the studies collected, 26 report the overall number of women in the study as well as the number experiencing any form of urinary incontinence (UI). Additionally, many studies provide data on two subtypes of UI: stress urinary incontinence (SUI) and urgency urinary incontinence (UUI). We present the data in Table A.1, which was not included in the original paper by Markland et al. (2018) [19]. Let y_{i0}, y_{i1}, and y_{i2} denote the number of women experiencing any UI, SUI and UUI in study i, respectively, and let n_i denote the total number of women in each study. While all studies provide counts for the total number of women experiencing any urinary incontinence, many studies do not report one or both subtype counts. We note that $y_{i0} \geq y_{i1}, y_{i2}$ and the counts for each subtype do not necessarily sum to the overall UI count when they are all reported, as they are not mutually exclusive. Therefore a model based on a multinomial distribution such as that used by Trikalinos et al. (2013) [18] would not be appropriate. Our goal is to estimate the population-averaged marginal prevalence of urinary incontinence (π_0), as well as that of each subtype (π_1, π_2).

Currently, standard practice would be to estimate each prevalence individually, using univariate random-effects models. However, it is reasonable to expect that the prevalences of the different subtypes of urinary incontinence outcomes will be correlated with one another, in addition to being correlated with the overall prevalence. This can allow us to use data from non-missing outcomes to address ORB and increase the precision of our pooled subtype estimates. Therefore, we first fit a Bayesian multivariate random effects model in order to

incorporate information about the correlations between the UI outcomes (π_0, π_1, π_2). We then compare these results to those found using a novel parameterization of the model that incorporates the natural constraint that $\pi_0 \geq \pi_1, \pi_2$. Finally, we compare the results found using both of these models to those found using separate univariate random effects models.

2.3 Methodology

2.3.1 A univariate random effects model

As mentioned previously, one way to model the overall and subtype prevalences of a condition, is to model each using separate univariate random effects models [41]. In using random effects models, we assume that the true prevalences vary across studies. Let π_{ij} be the probability of having the jth outcome in study $i \in \{1, ..., N\}$, where $j \in \{0, 1, ..., J\}$, and let n_i be the number of participants in study i. Here, π_{i0} refers to the overall prevalence in study i, while $\pi_{i1}, ..., \pi_{iJ}$ refer to the J subtype prevalences. Let S_i be the set of outcomes that are reported in study i, and $D_i = \{(y_{ij}, n_i), j \in S_i\}$ denote the available data from study i. Let $\phi(z)$ and $\Phi(\cdot)$ denote the probability density function and cumulative density function (CDF) of the standard normal distribution, respectively. If we use a probit link function to separately model the number of cases y_{ij} for each of the $J + 1$ total outcomes, we have

$$y_{ij} \sim \text{ Binomial}(n_i, \pi_{ij}), \Phi^{-1}(\pi_{ij}) = \mu_j + v_{ij}, v_{ij} \sim N(0, \sigma_j^2), \quad j \in S_i, i = 1, ..., N, \quad (2.1)$$

where μ_j is the fixed effect for each outcome, while v_{ij} is the random effect for outcome j within study i. Therefore, σ_j describes the between-study variability in outcome j. If we assume that conditional on the π_{ij}'s, the y_{ij}'s are independent, this gives us the following observed data likelihood function combining the $J + 1$ independent random effects models:

$$L_1 \propto \prod_{i=1}^{N} \prod_{j \in S_i} \int_{-\infty}^{\infty} [\Phi(\mu_j + \sigma_j z)]^{y_i} [1 - \Phi(\mu_j + \sigma_j z)]^{n_i - y_i} \phi(z) dz. \tag{2.2}$$

The population averaged prevalence for each outcome can be estimated as $\pi_j = E[\pi_j | \mu_j, \sigma_j] = \Phi\left(\frac{\mu_j}{\sqrt{1+\sigma_j^2}}\right)$ [30], $j \in \{0, 1, ..., J\}$.

2.3.2 A multivariate random effects model

We can extend this univariate approach to jointly estimate the prevalences for multiple outcomes, using a multivariate Bayesian random effects model. This incorporates the correlations between outcomes to improve estimation when missing data are present, otherwise known as borrowing strength. Zhang et al. (2014) [30] present a hierarchical Bayesian random effects model with a probit-link in the context of "arm-based" network meta-analysis. In "arm-based" NMA, the focus is on calculating the event probabilities for each treatment arm [30, 42, 43]. Therefore, we adapt this framework in order to estimate the event probabilities for our overall outcome and each of the subtypes, treating them as separate "arms."

We first let each y_{ij} be independently binomially distributed, conditional on π_{ij}. We then let the probit-transformed $(\pi_{i0}, ..., \pi_{iJ})^T$ follow a multivariate normal distribution:

$$\begin{aligned} y_{ij} &\sim \text{ Binomial}(n_i, \pi_{ij}) \\ [\Phi^{-1}(\pi_{i0}), \Phi^{-1}(\pi_{i1}), ..., \Phi^{-1}(\pi_{iJ})]^T &= (\mu_0 + v_{i0}, \mu_1 + v_{i1}, ..., \mu_J + v_{iJ})^T, \\ \mathbf{v}_i &= (v_{i0}, v_{i1} ..., v_{iJ})^T \sim MVN(\mathbf{0}, \Sigma_{J+1}), j \in S_i, \end{aligned} \tag{2.3}$$

where $\mathbf{\Sigma}_{J+1} = diag(\boldsymbol{\sigma})\mathbf{R}_{J+1}diag(\boldsymbol{\sigma})$. $\mathbf{R}_{J+1}$ is the within-study correlation matrix and the σ_j^2 terms capture the between study variation in each outcome. Let L_{ij} denote the conditional likelihood given v_{ij}, defined as

$$L_{ij}(y_{ij}; \mu_j, v_{ij}) = [\Phi(\mu_j + v_{ij})]^{y_{ij}} [1 - \Phi(\mu_j + v_{ij})]^{n_i - y_{ij}}. \tag{2.4}$$

Then the observed data likelihood function can be written as follows:

$$L_2 \propto \prod_{i=1}^{N} \int_{\mathbb{R}^J} \left(\prod_{j \in S_i} L_{ij}(y_{ij}; \mu_k, v_{ij}) \right) \frac{exp(-\frac{1}{2}\mathbf{v}_i^T \Sigma_{J+1}^{-1} \mathbf{v}_i)}{(2\pi)^{(J+1)/2} |\Sigma_{J+1}|^{1/2}} d\mathbf{v}_i, \tag{2.5}$$

where $\mathbb{R}^J$ is the J-dimensional real space. The population averaged prevalence for each outcome can be estimated using the same method used in the univariate model as $\pi_j = \Phi\left(\frac{\mu_j}{\sqrt{1+\sigma_j^2}}\right)$, $j \in \{0, 1, .., J\}$.

2.3.3 A new multivariate random effects model accounting for the natural constraint

While the above multivariate model allows us to incorporate information about the correlation between outcomes into our estimation of the population-averaged prevalences, it fails to account for the natural constraint in the data when we model an overall outcome along with the prevalence of different subtypes. We present a different parameterization of the previous model in order to account for this natural constraint: $y_{i0} \geq y_{i1}, y_{i2}, ..., y_{iJ}$, without requiring the subtypes be mutually exclusive.

First, let y_{i0} be binomially distributed with parameter π_{i0} as for the other two approaches. Then, let p_{ij} denote the proportion of cases in study i that fall into the subtype $j \in \{1, ..., J\}$. Therefore, y_{ij} is binomially distributed with denominator y_{i0} and probability p_{ij}. Let μ_0 denote the fixed effect for the overall outcome, and v_{i0} are the within-study random effects for this overall outcome. Let μ_j^* be the fixed effect corresponding to the proportion of outcomes that fall into subtype j and the v_{ij}^*'s be the corresponding random effects. We again use a probit link function to model π_{i0} and p_{ij}:

$$\begin{aligned} y_{i0} \sim\ & \text{Binomial}(n_i, \pi_{i0}), \Phi^{-1}(\pi_{i0}) = \mu_0 + v_{i0}, \\ y_{ij} \sim\ & \text{Binomial}(y_{i0}, p_{ij}), \Phi^{-1}(p_{ij}) = \mu_j^* + v_{ij}^*, j \in \{1, ..., J\}, j \in S_i, \\ & \mathbf{v^*}_i = (v_{i0}, v_{i1}^*, ..., v_{iJ}^*)^T \sim MVN(\mathbf{0}, \Sigma^*{}_{J+1}), \end{aligned} \tag{2.6}$$

where $\mathbf{\Sigma}^*_{J+1} = diag(\boldsymbol{\sigma^*})\mathbf{R}^*{}_{J+1}diag(\boldsymbol{\sigma^*})$. $\mathbf{R}^*{}_{J+1}$ is the within-study correlation matrix and the σ^{*2}_j terms capture the between study variation in the overall outcome and proportion of events that fall into each subtype. The key difference between Σ_{J+1} and $\Sigma^*{}_{J+1}$ is that the final J variance components $\sigma^{*2}_1, ..., \sigma^{*2}_J$ in $\Sigma^*{}_{J+1}$ describe the within-study random effects in the proportions of the overall outcome that fall into the individual subtypes $(v^*_{i1}, ..., v^*_{iJ})^T$, while $\sigma^2_1, ..., \sigma^2_j$ in Σ_{J+1} describe the within-study random effects in the prevalences of the individual subtypes $(v_{i1}, ..., v_{iJ})^T$. Furthermore, the correlation terms between subtypes can be interpreted as reflecting the correlation between the subtype event rates, conditional on the overall count in each study.

This results in the following observed data likelihood function:

$$L_3 \propto \prod_{i=1}^{N} \int_{\mathbb{R}^J} \left(\prod_{j \in S_i} L_{ij}(y_{ij}; \mu^*_k, v^*_{ij}) \right) \frac{exp(-\frac{1}{2}\mathbf{v}^*{}^T_i \Sigma^{-1}_{J+1} \mathbf{v}^*{}_i)}{(2\pi)^{(J+1)/2} |\Sigma^*{}_{J+1}|^{1/2}} d\mathbf{v}^*{}_i, \tag{2.7}$$

where the conditional likelihood L_{i0} is defined as

$$L_{i0}(y_{i0}; \mu_0, v_{i0}) = [\Phi(\mu_0 + v_{i0})]^{y_{i0}} [1 - \Phi(\mu_0 + v_{i0})]^{n_i - y_{i0}}, \tag{2.8}$$

and the conditional likelihood L_{ij} for $j \in \{1, ..., J\}$ is defined as

$$L_{ij}(y_{ij}; \mu^*_j, v^*_{ij}) = \left[\Phi(\mu^*_j + v^*_{ij})\right]^{y_{ij}} \left[1 - \Phi(\mu^*_j + v^*_{ij})\right]^{y_{i0} - y_{ij}}. \tag{2.9}$$

We can then estimate π_0 by:

$$\pi_0 = E[\pi_{i0} | \mu_0, \sigma_0] = \int_{-\infty}^{\infty} \Phi(\mu_0 + \sigma_0 z)\phi(z)dz = \Phi\left(\mu_0 / \sqrt{1 + \sigma^2_0}\right), \tag{2.10}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function, $Z \sim N(0, 1)$, and $\phi(\cdot)$ is the density of the standard normal distribution. Let (X, Y) be a standard bivariate normal with covariance $Cov_{X,Y} = \frac{1}{\sqrt{1+\sigma^2_0}\sqrt{1+\sigma^{*2}_j}} \Sigma^*{}_{1,j+1}$. As shown in the Supplementary

Materials, we can estimate $\pi_j = E[\pi_{i0} p_{ij} | \mu_0, \mu_j^*, \sigma_0, \sigma_j^*]$, $j \in \{1, .., J\}$ using:

$$\begin{aligned} \pi_j &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi(\mu_0 + \sigma_0 z_0) \Phi(\mu_j^* + \sigma_j^* z_j) \phi(z_0, z_j) dz_0 dz_j \\ &= P\left(X < \frac{\mu_0}{\sqrt{1+\sigma_0^2}}, Y < \frac{\mu_j^*}{\sqrt{1+\sigma_j^{*2}}} \right). \end{aligned} \tag{2.11}$$

This new parameterization of the multivariate random effects model truncates the density of each study-level subtype prevalence at the current estimate for the overall prevalence. We hypothesized that this would decrease bias and increase precision when estimating the population-averaged subtype prevalences.

2.3.4 Prior specifications

We use Markov chain Monte Carlo (MCMC) methods to obtain Bayesian posterior estimates for each π_j and ${\Sigma^*}_{J+1}$, with $N(0, 1000)$ priors for each μ_j. For the univariate models, we put a $Unif(0, 10)$ prior on each σ_j. The inverse-gamma(ϵ, ϵ), where ϵ is small, has previously been a popular choice of prior for σ_j^2, as it is conditionally conjugate [44]. However, the results can be sensitive to the choice of ϵ, particularly for small σ [44]. We use the same choice of priors with a spherical decomposition for both ${\Sigma^*}_{J+1}$ and Σ_{J+1}. The commonly used conjugate inverse-Wishart prior for the precision matrix of multivariate normal random vectors can result in inflated estimates of the variances and shrinkage of the correlations towards zero, particularly when the true variances are small [29, 45]. Thus, we use a separation strategy [46] in order to specify the priors on Σ_{J+1} and Σ_{J+1}^*, which involves modeling the variance and correlation components separately. Specifically, in this case we use the spherical decomposition described by Lu and Ades (2009) [47] and Wei and Higgens (2013) [29] and implemented in the "pcnetmeta" R package [48], with $Unif(0, \pi)$ priors on the coordinate parameters [47].

2.4 Missing at random (MAR) assumption

While these methods do not require that all studies measure each outcome, they do require the assumption that the data are missing at random (MAR). That is, the probability of an outcome being measured for any given study may depend only the observed outcomes for that study. Since we are currently assuming that the overall outcome is observed for all studies, when we simulate data that are MAR, we let the probability of being missing depend only on the overall prevalence.

To illustrate, let $m_{ij} = 1$ if the number of events y_{ij} for the jth outcome of the ith study is not reported. We can then model the probability that $m_{ij} = 1$ by:

$$\text{logit}\left(P(m_{ij} = 1)\right) = \alpha_{0j} + \alpha_{1j} f(\pi_{i0}) + \alpha_{2j} g(\pi_{ij}), \tag{2.12}$$

where $f(\cdot)$ and $g(\cdot)$ can be any functions. The specific functions we used in our simulation studies are described in Section 2.5.1. The likelihood function modeling the missing data is $L_m \propto \prod_{i=1}^{N} \prod_{j=1}^{J} P(m_{ij} = 1)^{m_{ij}} [1 - P(m_{ij} = 1)]^{1-m_{ij}}$. Multiplying the likelihood for the observed data as in equation (2.5) or (2.7) with L_m, we obtain the total likelihood function which incorporates the missing mechanism. If the data are missing completely at random (MCAR), then $\alpha_{1j} = \alpha_{2j} = 0$, and the probability of the j^{th} outcome being missing is some constant determined by α_{0j}. If the data are missing at random, then $\alpha_{2j} = 0$ and the probability of being missing only depends on the underlying event rates for the data that are fully observed. Finally, if $\alpha_{2j} \neq 0$, then the data are missing not at random (MNAR), since the probability of an outcome being missing depends directly on the underlying event rate. We can evaluate the impact of the missing data assumptions by simulating patterns of missingness that correspond to different values of α_{1j} and α_{2j}.

2.5 Simulation Studies

2.5.1 Methods of simulation

In order to compare the performance of the univariate models, the standard multivariate model, and the new multivariate model, we performed two main sets of simulations: one with data for the two subtypes under MCAR and one under MAR. For each setting, we simulated 2000 data sets containing 30 studies. Each data set contains the overall count for each study (y_{i0}), the counts for two subtypes (y_{i1}, y_{i2}), and the overall number of participants in each study (n_i). We let y_{i0} be distributed binomial with denominator n_i and success probability π_{i0}, and y_{i1} and y_{i2} each be distributed binomial with denominator y_{i0} and success probabilities p_{i1} and p_{i2}, respectively. The $\left(\Phi^{-1}(\pi_{i0}), \Phi^{-1}(p_{i1}), \Phi^{-1}(p_{i2})\right)$ are distributed as a multivariate normal with different variances. For simplicity, we let all of the pairwise correlations be identical and equal to ρ. The overall count y_{i0} is observed in all studies, while the mean probabilities of missing each subtype across all studies, $\bar{m}_{\cdot 1}$ and $\bar{m}_{\cdot 2}$ are 0.5. Each condition is repeated twice, with n_i, the sample size for each study, equal to 100 or 500. All models are fit using JAGS version 4.3.0 [49], run using R version 3.3.3 [50] and packages "rjags" [51] and "coda" [52], and consist of 2 independent chains with 20,000 samples each. We also use a burn-in period of 5,000 samples and a thinning interval of 2. The 2,000 simulations for each condition were run in parallel using the Minnesota Supercomputing Institute (MSI) resources.

We first simulated data for 30 studies (N = 30), where observations for the two subtypes $(y_{i1}; y_{i2})$ were all MCAR with probability 0.5. We set the pairwise correlations between $\Phi^{-1}(\pi_{i0})$, $\Phi^{-1}(p_{i1})$, $\Phi^{-1}(p_{i2})$ to all be equal to ρ, which had possible values of $(0, 0.4, 0.8)$. The $(\sigma_0, \sigma_1, \sigma_2)$ were set to be either (0.5, 0.5, 0.5) or (0.5, 1, 1). The μ_0, μ_1 and μ_2 are set such that the population-averaged prevalence (π_0, π_1, π_2) were equal to (0.3, 0.15, 0.05), respectively. This corresponded to 6 different conditions, which we repeated with $n_i = (100, 500)$ for each of the 30 studies, for a total of 12 conditions. To investigate the

performance of proposed methods under small number of studies, we also included a set of simulations when the number of studies $N = 10$ and the sample size per study $n_i = 100$, giving an additional 6 scenarios.

Second, we repeat each of the MCAR conditions described above, but with y_{i1} and y_{i2} under MAR and the marginal probability of being missing across all studies $P(m_{ij} = 1) = 0.5$ for $j = 1, 2$. We let the probability of missingness for the jth subtype of the ith study, $P(m_{ij} = 1)$ depend on π_{i0}, the observed overall study-specific prevalence. For simplicity, we assume $P(m_{i1} = 1) = P(m_{i2} = 1)$, since these only depend on the overall study-specific prevalence. From equation (2.12), we let $\alpha_{0j} = \text{logit}(0.5)$, $\alpha_{1j} = 3$, $j = 1, 2$, and $f(\pi_{i0}) = \text{logit}(\pi_{i0}) - \text{logit}(\bar{\pi}_{i0})$ such that $P(m_{ij} = 1)$ is inversely proportional to the difference in the logit of overall study-specific prevalence and its mean, i.e.,

$$P(m_{ij} = 1) = \text{logit}^{-1}(\text{logit}(0.5) - 3(\text{logit}(\pi_{i0}) - \text{logit}(\bar{\pi}_{.0})). \tag{2.13}$$

Therefore, studies with smaller prevalence will be more likely to be missing, leading to ORB as described in Section 2.1.

Finally, we also repeated each of the 6 conditions where $n_i = 100$ for the case were there were no missing data. This included separate cases where each data set included 30 studies or 10 studies. Overall, this corresponded to 18 MCAR and MAR conditions and 12 no missing data conditions, for a total of 48 conditions.

Table 2.1: **(N = 30, n_i = 100, MCAR)** Bias, 95% credible interval width (CIW) and coverage probability (Cov.) for univariate model, original multivariate model, and new parameterization, across 2000 simulations containing 30 studies, where $n_i = 100$, the true prevalences are (0.3, 0.15, 0.05), and data are missing completely at random (MCAR).

		Overall			**Subtype 1**			**Subtype 2**		
		Bias	CIW	Cov.	Bias	CIW	Cov.	Bias	CIW	Cov.
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \rho = 0$	Univariate	0.003	0.121	0.95	0.009	0.127	0.96	0.008	0.078	0.96
	Original	0.003	0.122	0.96	0.007	0.102	0.95	0.008	0.071	0.95
	New Param.	0.004	0.124	0.96	0.002	0.097	0.95	0.005	0.061	0.95
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \rho = 0.4$	Univariate	0.003	0.121	0.95	0.010	0.143	0.97	0.011	0.094	0.97
	Original	0.003	0.122	0.96	0.007	0.109	0.96	0.009	0.076	0.95
	New Param.	0.004	0.124	0.96	0.003	0.103	0.95	0.006	0.066	0.94
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \rho = 0.8$	Univariate	0.002	0.121	0.96	0.012	0.158	0.97	0.013	0.105	0.96
	Original	0.002	0.121	0.95	0.006	0.114	0.96	0.007	0.074	0.96
	New Param.	0.003	0.122	0.96	0.004	0.107	0.96	0.005	0.066	0.95
$\boldsymbol{\sigma} = (0.5, 1, 1), \rho = 0$	Univariate	0.002	0.121	0.95	0.013	0.162	0.96	0.018	0.132	0.97
	Original	0.003	0.123	0.96	0.012	0.146	0.95	0.021	0.133	0.95
	New Param.	0.004	0.124	0.96	0.001	0.121	0.96	0.010	0.089	0.95
$\boldsymbol{\sigma} = (0.5, 1, 1), \rho = 0.4$	Univariate	0.002	0.121	0.96	0.015	0.183	0.97	0.023	0.153	0.97
	Original	0.003	0.123	0.96	0.012	0.152	0.96	0.022	0.137	0.95
	New Param.	0.004	0.124	0.96	0.003	0.129	0.96	0.011	0.097	0.95
$\boldsymbol{\sigma} = (0.5, 1, 1), \rho = 0.8$	Univariate	0.002	0.121	0.96	0.018	0.203	0.97	0.025	0.166	0.96
	Original	0.002	0.121	0.96	0.010	0.148	0.96	0.018	0.123	0.96
	New Param.	0.003	0.122	0.96	0.004	0.132	0.95	0.011	0.096	0.95

2.5.2 Simulation results

Table 2.1 summarizes the results where N = 100 and approximately 50% of the subtype data are MCAR across studies. All three models gave similar results for the fully observed overall outcome, with the univariate model having slightly less bias and shorter credible intervals than the two multivariate models. However, using the new parameterization reduced both

bias and 95% credible interval width (CIW) for the two subtypes under all conditions, outperforming both the univariate and original multivariate models. This reduction in bias and CIW for the two subtypes became larger as both ρ and the subtype variance increased. The original multivariate model still reduced bias and CIW over the univariate approach, except when $\rho = 0$, where it corresponded to a larger bias, CIW, or both. We observe qualitatively similar results for each condition when N = 500, as well as the conditions where each data set contained only 10 studies, which are given in Tables A.2 and A.4 in the Supplementary Materials, respectively.

Table 2.2: **(N = 30, n_i = 100, MAR)** Bias, 95% credible interval width (CIW) and coverage probability (Cov.) for for univariate model, original multivariate model, and new parameterization, across 2000 simulations containing 30 studies, where $n_i = 100$, the true prevalences are (0.3, 0.15, 0.05), and data are missing at random (MAR).

		Overall			**Subtype 1**			**Subtype 2**		
		Bias	CIW	Cov.	Bias	CIW	Cov.	Bias	CIW	Cov.
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \rho = 0$	Univariate	0.003	0.121	0.95	0.064	0.127	0.38	0.027	0.081	0.65
	Original	0.003	0.123	0.96	0.017	0.121	0.94	0.014	0.089	0.93
	New Param.	0.004	0.124	0.96	0.002	0.092	0.96	0.005	0.057	0.96
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \rho = 0.4$	Univariate	0.003	0.121	0.95	0.078	0.150	0.35	0.037	0.102	0.56
	Original	0.003	0.122	0.96	0.014	0.116	0.94	0.013	0.080	0.94
	New Param.	0.004	0.124	0.96	0.005	0.099	0.95	0.007	0.060	0.96
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \rho = 0.8$	Univariate	0.003	0.121	0.96	0.092	0.161	0.27	0.045	0.108	0.42
	Original	0.003	0.121	0.95	0.013	0.113	0.94	0.010	0.070	0.95
	New Param.	0.003	0.122	0.96	0.008	0.104	0.95	0.007	0.058	0.94
$\boldsymbol{\sigma} = (0.5, 1, 1), \rho = 0$	Univariate	0.003	0.121	0.95	0.068	0.189	0.65	0.040	0.159	0.83
	Original	0.004	0.124	0.96	0.032	0.195	0.95	0.038	0.185	0.93
	New Param.	0.004	0.124	0.96	0.002	0.114	0.96	0.011	0.084	0.95
$\boldsymbol{\sigma} = (0.5, 1, 1), \rho = 0.4$	Univariate	0.003	0.121	0.96	0.090	0.202	0.50	0.052	0.171	0.67
	Original	0.004	0.123	0.96	0.025	0.172	0.95	0.029	0.152	0.93
	New Param.	0.005	0.124	0.97	0.007	0.123	0.96	0.013	0.087	0.95
$\boldsymbol{\sigma} = (0.5, 1, 1), \rho = 0.8$	Univariate	0.003	0.121	0.96	0.112	0.210	0.34	0.063	0.173	0.55
	Original	0.003	0.121	0.96	0.016	0.141	0.93	0.016	0.104	0.95
	New Param.	0.004	0.123	0.96	0.011	0.128	0.94	0.011	0.084	0.94

The reduction in bias for the two subtypes for the two multivariate parameterizations increased when the data were MAR, as shown in Table 2.2 (N = 100) and Table A.3 (N = 500). This includes the conditions in the MCAR scenarios (where $\rho = 0$) where the original multivariate model sometimes had larger bias than the univariate models for the subtype outcomes. In the MAR case where $\rho = 0, \sigma = (0.5, 1, 1)$, and N = 100, using the original

multivariate parameterization reduced bias by 52.9% and 5.0% for subtypes 1 and 2, respectively, while using the new parameterization reduced it by 97.1% and 72.5%, respectively, when compared to the univariate models. As expected, the coverage probabilities for the univariate models were quite low when the data were MAR (as low as 23.4% for the N = 500, $\sigma = (0.5, 0.5, 0.5)$, $\rho = 0.8$ scenario). We again observed similar results when N = 10 (Table A.5).

The results for the cases where there were no missing data are presented in Tables A.6 and A.7 in the Supplementary Materials for N = 30 and N = 10, respectively. Here, the original multivariate model provided little to no benefit over using separate univariate models. Under many conditions, this model had larger bias and wider credible intervals on average than the separate univariate models. We hypothesize that without any missing data, the original multivariate model's ability to borrow information across outcomes may not make up for the additional model complexity. These results are also consistent with previous literature [16, 31]. However, the new parameterization still had reduced bias and credible interval widths even without missing data. We hypothesize that this is due to the reduction in density where the subtypes would have greater prevalence than the overall outcome. Furthermore, in the case where the subtype outcomes are distributed as binomial with the denominator equal to the overall count, the probit (or logit) transformed prevalences likely do not follow a normal distribution. This could further explain the reduction in bias found with the new parameterization.

In summary, using the two multivariate parameterizations did not improve performance when focusing solely on estimating the overall prevalence. However, these models improved estimation of the prevalences of the two subtypes over using separate univariate models, particularly under the MAR conditions, with the new parameterization outperforming the original multivariate model under all conditions.

2.6 The Case Study

2.6.1 Results under missing at random assumption

Table 2.3a presents estimates of the overall urinary incontinence (UI) prevalence and that of each subtype (SUI, UUI) found using separate univariate models, the multivariate model, and the new parameterization accounting for the subtype constraint. Each model was fit using 3 independent chains with 250,000 samples each and a burn-in period of 5,000 samples. We used a substantially larger number of iterations for the case study in order to generate smooth plots of the estimated density of the posterior predictive distribution. We also report the posterior mean and standard deviation of the between study variances and the correlations between outcomes in Table 2.3b. We expect the estimates of the covariance matrices to differ between the two parameterizations, since the random effects for the subtypes in the new parameterization refer to the variability in the proportion of overall cases that fall into that subtype, rather than the variability in study specific prevalence.

Table 2.3: **Case Study Results**

Model	UI	CIW	SUI	CIW	UUI	CIW
Univariate	0.274 (0.024)	0.096	0.127 (0.022)	0.088	0.066 (0.021)	0.082
Multivariate	0.275 (0.025)	0.098	0.128 (0.022)	0.088	0.064 (0.019)	0.072
New Parameterization	0.275 (0.025)	0.099	0.123 (0.02)	0.078	0.061 (0.016)	0.064

(a) Posterior mean (SD) of marginal prevalences for overall outcome (UI) and two subtypes (SUI, UUI) with corresponding 95% credible interval width (CIW)

	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$
Univariate	0.383 (0.060)	0.435 (0.089)	0.607 (0.131)
Multivariate	0.394 (0.062)	0.457 (0.093)	0.6 (0.121)
	$\hat{\sigma}_1^*$	$\hat{\sigma}_2^*$	$\hat{\sigma}_3^*$
New Param.	0.399 (0.064)	0.755 (0.160)	0.683 (0.157)

	$\hat{\rho}_{12}$	$\hat{\rho}_{13}$	$\hat{\rho}_{23}$
Multivariate	0.462 (0.204)	0.681 (0.158)	0.395 (0.220)
	$\hat{\rho}_{12}^*$	$\hat{\rho}_{13}^*$	$\hat{\rho}_{23}^*$
New Param.	-0.115 (0.249)	0.364 (0.233)	0.103 (0.270)

(b) Posterior mean (SD) of components in estimated covariance matrices for original multivariate model and new parameterization ($\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}}^*$)

All three models gave similar results for the overall prevalence of UI. The estimates of the SUI prevalence differed slightly across the three models, with the new parameterization giving the smallest estimate. The proportion falling into the SUI subtype also had low correlation with the overall outcome under the new parameterization (-0.115) and larger variance (0.755). The 95% credible interval was wider for the original parameterization, than the univariate model, similar to the $\sigma = (0.5, 1, 1)$ and $\rho = 0$ scenarios from Section

2.5.2. However, the new parameterization still reduced 95% credible interval width by 11.4% compared to the univariate model.

The estimated marginal prevalence of UUI was highest under the univariate model, and lowest under the new parameterization. The proportion falling into the UUI category had a higher estimated correlation with the overall outcome under the new parameterization (0.364) and slightly lower estimated variance (0.600) when compared to the SUI outcome. Jointly modeling the outcomes resulted in a 12.2% and 22.0% reduction in 95% credible interval width for the original and new multivariate parameterizations, respectively. Based off of the original model, the estimated unconditional correlation between the SUI and UUI subtypes was relatively low (0.395). As estimated by the new parameterization, the estimated correlation conditional on the overall count was even smaller (0.103).

Figure 2.1 presents forest plots for each outcome, showing the shrunken estimates for the study-level prevalences under the three different models. The estimates and credible intervals are similar across all three methods for the overall prevalence of UI, as well for the subtypes in studies where the given outcome is observed. While the univariate model cannot estimate study level prevalence for unobserved subtypes, the two multivariate parameterizations can do so using information from the correlations and observed outcomes. The new parameterization gave narrower credible intervals than the original multivariate model when estimating subtype prevalence corresponding to the unobserved outcomes. Similarly, Figure A.1 presents posterior density plots for each outcome, by model. This illustrates the reduction in density at larger values when using the new parameterization.

Figure 2.1: **Forest Plot of Study Level Estimates** Posterior mean and 95% credible interval of marginal and study level prevalences for each of the three outcomes across 26 studies

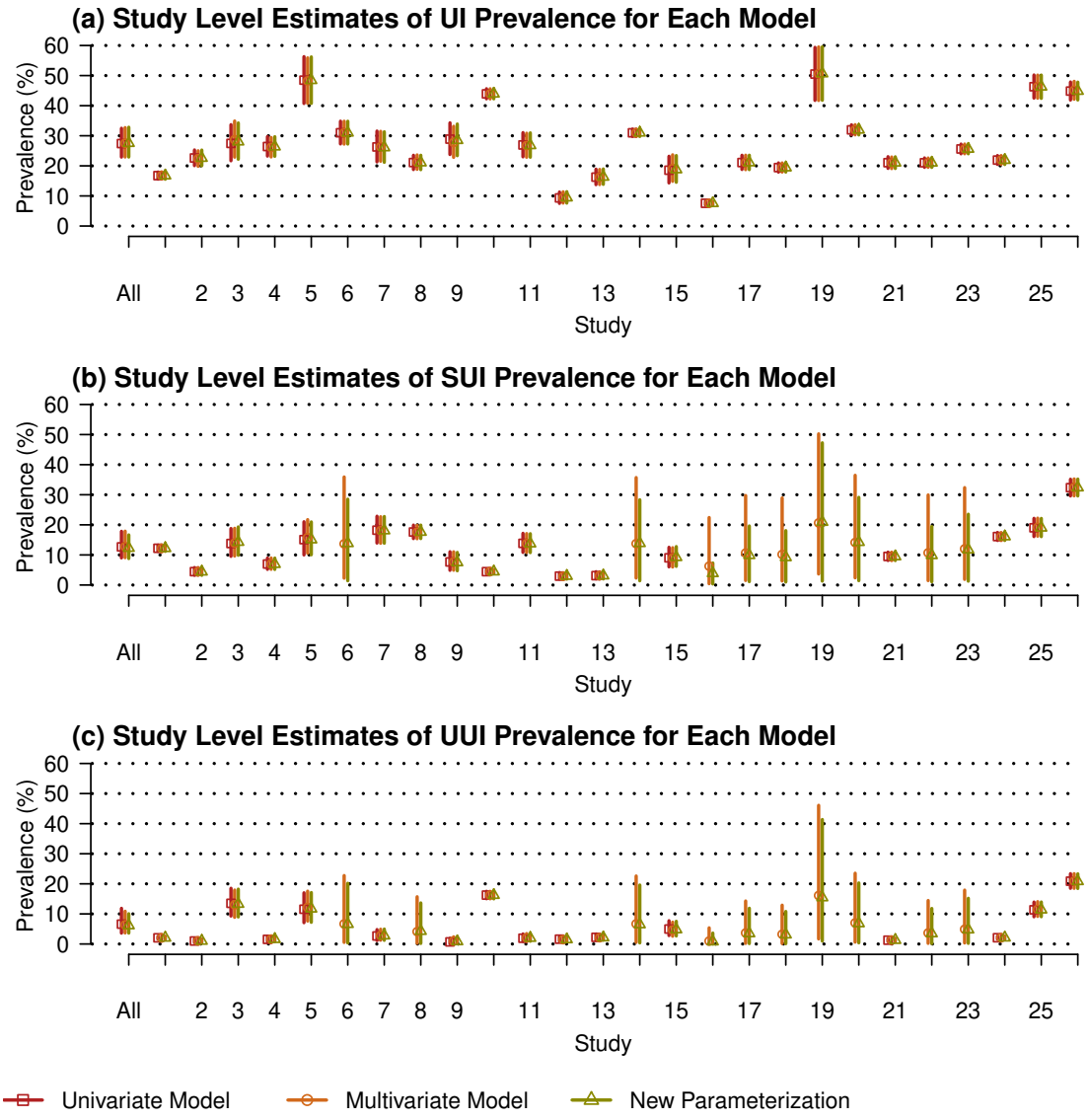
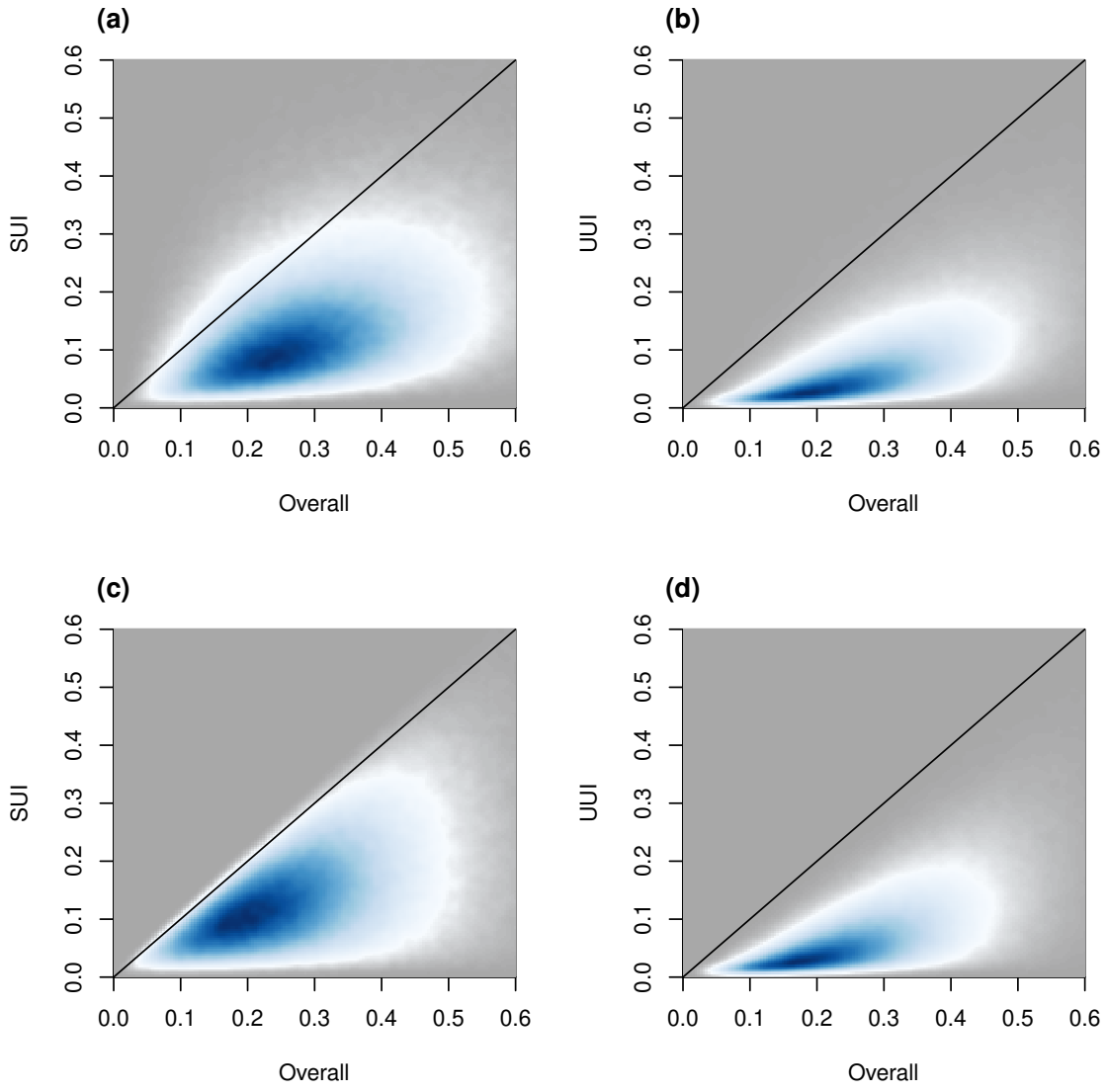
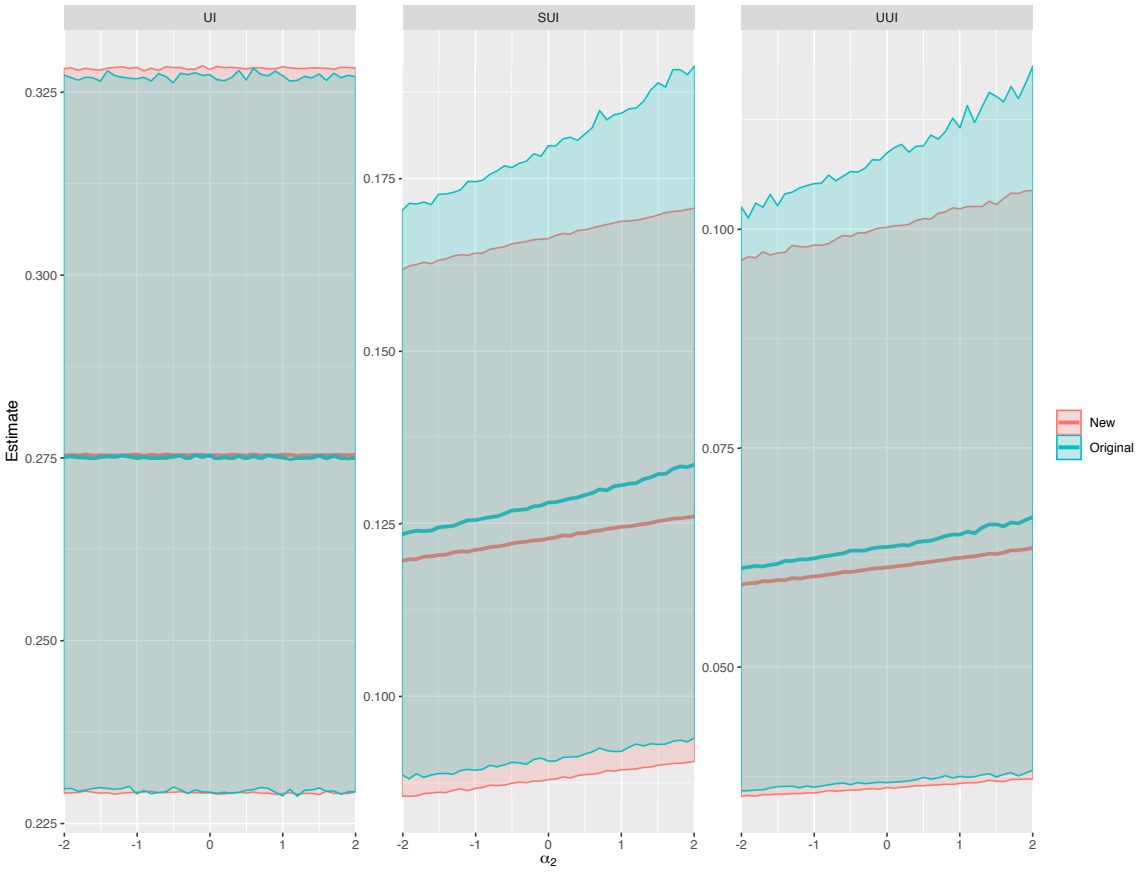


Figure 2.2: **Bivariate Density Plots for Predicted Prevalences in New Study** (a) Overall and SUI posterior predicted prevalences for new study based on original multivariate model results, (b) Overall and UUI with original multivariate model, (c) Overall and SUI with new parameterization, (d) Overall and UUI with new parameterization



We further investigated this reduction in joint density by the new parameterization using the posterior predictive bivariate density plots, as shown in Figure 2.2. We estimated these posterior predictive distributions by drawing a random sample representing the prevalences of a future study from the joint posterior distributions of parameters at each of the 750,000 total iterations. Of the samples generated according to the results from the original multivariate model, 64 samples and 1 sample gave estimates where the overall UI prevalence was smaller than the corresponding estimates for the SUI and UUI prevalences, respectively. The new parameterization on the other hand specifically prevents this from occurring, which explains the reduction in density between (a) and (c) in Figure 2.2. Finally, we tested the sensitivity to the choice of the separation strategy prior on the covariance matrix by rerunning the two multivariate models using an inverse-Wishart prior. The results are included in Table A.8 in the Supplementary Materials, and are similar to those found using the previous choice of prior.

Figure 2.3: **Sensitivity Analysis** Posterior mean and 95% credible intervals for UI, SUI, and UUI marginal prevalence across values of α_2



2.6.2 Sensitivity analysis results under MNAR

When fitting each model, we assume that the probability of each study missing a subtype outcome does not directly depend on the underlying prevalence. As we cannot directly test whether the data are MNAR, we conduct a sensitivity analysis assuming several different patterns of missingness to evaluate the impact of different degrees of MAR violations. We assume that $m_{ij} \sim Ber(q_{ij})$, where q_{ij} is the probability of subtype j for study i being

missing. Specifically, we specify a logistic model for q_{ij} as $\text{logit}\,(q_{ij}) = \alpha_{0j} + \alpha_{2j}\pi_{ij}$, where α_{2j} is not identifiable. Instead, we specify values for $\alpha_{2j} \in [-2, 2]$, with increments of 0.1 and observe how the estimates of the marginal prevalences π_0, π_1, π_2 vary when (2.12) is incorporated into the likelihoods for the original multivariate model (2.5) and new parameterization (2.7). For simplicity, we let $\alpha_{21} = \alpha_{22} = \alpha_2$ for each $\alpha_{2j} \in [-2, 2]$.

Figure 2.3 presents the posterior means and 95% credible intervals of π_0, π_1, and π_2 for the two models under each value of $\alpha_2 \in [-2, 2]$. As expected, the posterior mean for the overall UI prevalence (π_0) remained approximately constant across values of α_2, since it is fully observed. However, the posterior means for the two subtype prevalences (π_1, π_2) generally increased with α_2. Larger values of α_2 corresponded to studies with larger π_{ij} missing with higher probability, thus the estimated prevalence became larger in order to incorporate this fact. The posterior mean of π_1 ranged from (0.123, 0.133) and from (0.120, 0.126) for the original and new parameterizations, respectively, while the posterior mean of π_2 ranges from (0.061, 0.067) and from (0.059, 0.064). Therefore, violating the MAR assumption to this extent led to a slightly larger difference in estimates for the original model than for the new parameterization. We also note that the upper limit of the 95% credible interval from the new parameterization is consistently lower than that from the original model for the SUI and UUI panels of Figure 2.3.

In general, the mechanism of MNAR is unobservable. The sensitivity analysis above was intended to examine the risk of bias under a few MNAR mechanisms. Missingness may depend on other unobserved characteristics of the study population and even if missingness is only related to subtype prevalences, the dependency may differ from what we considered.

2.7 Discussion

Jointly modeling the overall and each subtype prevalence using multivariate random effects models generally improved both bias and precision for the marginal subtype estimates by

"borrowing strength" across outcomes. Incorporating the natural subtype constraint improved estimation further, likely by eliminating density in regions where the subtype prevalence would be larger than the overall prevalence for an individual study. This occurred even under the few conditions where the original multivariate parameterization failed to improve estimation, including when there were no missing data. Using the original multivariate model was the least beneficial when correlations between outcomes were low, variances were large, and the data were MCAR. We hypothesize that the multivariate models cannot borrow much information when $\rho = 0$. In this case, the multivariate model required estimating far more parameters than the univariate models, thus increasing variability. However, the new parameterization improved estimation under these conditions by taking into account the natural constraint in the data. Both models improved bias when the data were MAR and correlations were large, with the new parameterization still outperforming the original multivariate model.

This behavior was reflected in the results of the case study. The SUI subtype had low correlation with the overall outcome under the new parameterization and slightly larger variance. This likely led to the original multivariate model giving a wider 95% credible interval than the univariate model, while the new parameterization having a slightly narrower credible interval. The UUI subtype had a slightly higher correlation with the overall outcome and slightly lower variance. In this case, the original multivariate model had a slightly shorter 95% credible interval than the univariate model, consistent with the results of the simulations. We were also able to directly observe the reduction in density in the posterior predictive distributions for a single study associated with using the new parameterization, as shown in Figure 2.2. Finally, using the original and new multivariate parameterizations to jointly model the case study outcomes allowed us to estimate the correlation between subtypes both unconditionally and conditioned on the overall outcome.

Because the simulation conditions were generated under the new parameterization in order to ensure the subtype counts did not exceed the overall counts, the correlations used to

generate the data correspond to the overall outcome and proportions falling into each subtype. Large positive or negative correlations between the two subtypes would imply that the subtypes had a high co-morbidity or tended toward mutual exclusivity, respectively. Therefore, we hypothesize that using the new parameterization would be the most advantageous when a disease or set of outcomes are suspected to have either of these characteristics. However, if the subtypes were known to be mutually exclusive, a multinomial model would be more appropriate. We also caution that the methods may not be appropriate in the case of very rare diseases, resulting in very sparse counts, as the resulting estimates may not be stable.

Because we based the simulated marginal prevalences on those observed in the case study, all simulations were conducted using $(\pi_0, \pi_1, \pi_2) = (0.3, 0.15, 0.05)$. Therefore, the elimination of density by the new parameterization excluded all values above approximately 0.3, omitting about 70% of the original space. Furthermore, we hypothesize that this reduction in density may have the most impact in cases where at least one subtype has a high prevalence relative to the overall outcome. In this situation, the original multivariate model would likely have a greater number of samples where the subtype prevalence estimate was incorrectly higher than the overall prevalence. While we assume in this paper that the overall outcome is fully observed, these methods could be used when some studies do not report the overall count, such as in the food allergy [13] and autism spectrum disorder (ASD) [14] meta-analyses mentioned in Section 2.1. While both multivariate methods can be implemented with multiple subtypes, the estimates may become unstable in higher dimensions depending on the amount of missing data, variances, and correlations between outcomes. More work would be needed to evaluate these models' performance in higher dimensional scenarios.

As mentioned in Section 2.12, both multivariate methods require that any missing data be MAR or MCAR. Our analysis of the case study data included a sensitivity analysis of a specific MNAR mechanism, but the true missingness mechanism remains unknowable. However, we hypothesize that it may be less common for the type of prevalence data we

have described to be MNAR. If the data were MNAR, the probability of a given subtype being missing would directly depend on the underlying prevalence of each study. One such scenario would be given by investigators choosing not to report the prevalence of a subtype with a low count. We view this as unlikely, since if a study aiming to estimate the prevalence of a subtype encountered a lower frequency than expected, the outcome would still be of interest. Furthermore, while we have described in Section 2.1 the possible situation of studies measuring a lower overall prevalence being less likely to measure the lower frequency subtypes, this would be a case of the data being MAR, as the probability of being missing would depend on the overall count.

As discussed in Section 2.3.4, we use the separation strategy initially proposed by Barnard et al. (2000) [46] for the priors on the covariance matrices in the two multivariate models. The separation strategy has been shown to improve estimation of the variance and covariance terms over the use of an inverse-Wishart conjugate prior by adding more flexibility. However, this method also greatly increased computational time, as the multivariate models took up to a few hours to fit, depending on the number of iterations used. Using an HMC sampler such as STAN may decrease computational time over JAGS (a Gibbs based sampler) [45].

To the best of our knowledge, there is only one other existing multivariate meta-analysis of multivariate prevalence data [10]. This serves as a case study illustrating the potential improvement in estimation by jointly modeling multivariate prevalence data, particularly when incorporating additional natural constraints into the model parameterization. The methods used in this analysis allowed us to better compare prevalence rates of specific lower urinary tract symptom types across different occupation types in working women. These comparisons have informed future research on a wider range of lower urinary tract symptoms, in addition to urinary incontinence.

Chapter 3

Estimating the Reference Range from a Meta-Analysis

3.1 Introduction

The number of published meta-analyses has increased sharply over the past several decades [1, 37]. While most meta-analyses aim to provide a more precise estimate of the effect of a treatment or a risk factor's association with a disease [1], the literature has many examples of meta-analyses of normative data [7–9, 12, 53–59]. These studies generally aim to establish "typical" or "normal" values for a measurement or outcome using "healthy" populations from multiple studies, to serve as a reference. However, most often these meta-analysis studies report the pooled mean as the "reference value," which has limited interpretability when determining whether a measurement is "normal". Although Bohannon [53] noted that measurements lying outside the confidence interval for the pooled mean could be considered above or below "average", a reference range would be more useful in determining whether an observed measurement was within the range of values measured on healthy individuals. Horn et al. [21] define a reference range or interval as "a set of values within which some percentage, 95% for example, of the values of a particular analyte in a healthy population

would fall." In a meta-analysis, this requires accounting for the natural variability in the healthy population as reflected by variation both within and between studies.

Several medical systematic reviews have estimated and reported reference ranges using a meta-analysis of healthy individuals from multiple studies [7, 9, 56, 58, 60–63]. However, some of these studies have used the confidence interval for the pooled mean as the "reference range" [7, 61, 62], which reflects uncertainty in the estimated mean, not natural variation in the population. Venner et al. [63] used the measurement ranges reported in each study when available to construct reference ranges based on the overall minimum and maximum values across studies. While this better reflects natural variation across healthy individuals than the confidence interval for the pooled mean, only four out of the twelve studies included in the meta-analysis reported ranges, and this method requires setting the desired percentage of individuals captured in the reference range to 100%.

However, several studies estimate reference ranges containing a specified proportion of measurements from a healthy target population based on the observed mean, standard deviation, and sample size from each study [9, 56, 58, 60]. Conceição et al. [9] use a method similar to the empirical approach proposed later in this paper in order to estimate normal ranges for how accurately healthy participants perceive whether they are oriented vertically in space. Wyman et al. [58] use the fixed effects model by Laird and Mosteller [64] in order to establish normative ranges for non-invasive bladder function measurements in healthy women. Németh et al. [56] estimate a reference range for normal concentrations of asymmetric dimethylarginine in the plasma of healthy individuals, though their method for estimating the marginal reference range across all studies is not clear. Finally, Khoshdel et al. [60] simulate individual patient data based on the summary statistics from each study, then use fractional polynomials to estimate age-specific reference ranges for pulse wave velocity.

It is unknown what proportion of measurements from the "true" overall populations these reference ranges capture. Currently, the literature gives no guidance on how to approach

the question of estimating reference ranges based on meta-analyses. Several authors have recently advocated reporting prediction intervals for a new study [65–68], but they have not addressed prediction for an individual. To the best of our knowledge, the present paper is the first to propose methods for estimating reference ranges based on meta-analyses.

The first two proposed methods build on the commonly used random effects model. Section 3.2 motivates this choice and introduces notation. Section 3.3 proposes three approaches for estimating reference ranges. The first uses results from a frequentist method such as restricted maximum likelihood estimation (REML), while the second is a Bayesian approach using a posterior predictive interval. The final method is an empirical approach similar to that used by Conceição et al. [9]. We call these the frequentist, Bayesian, and empirical approaches. All three of these approaches use the means, standard deviations, and sample sizes reported in each study and do not require individual patient data. Section 3.4 presents simulation studies illustrating the performance of the three approaches, which we then apply to two examples in Section 3.5. Finally, Section 3.6 discusses some of the key distributional assumptions required by these approaches and potential areas of future work.

3.2 Random effects model

3.2.1 Choice of model

Three models are commonly used in meta-analysis: the common effect, random effects, and fixed effects models. The common effect model assumes the underlying true mean or effect is the same in each study and that variation between studies in estimated mean arises purely from sampling variation [69, 70]. This is often called a fixed effect model, which is easily confused with the fixed effects model of Laird and Mosteller [64]. We follow Bender et al. [69] by using the term "common effect model". This model imposes a strong assumption that each study population has the same underlying true mean, which may not be appropriate

for situations in which the observed means from the studies differ for reasons other than sampling variability [71]. When there is considerable heterogeneity between studies, as is often the case in meta-analyses of continuous outcomes [72], it may instead be desirable to assume that each study has a different true mean and that these means are drawn from separate distributions, as in the fixed effects model of Laird and Mosteller [64]. However, because this model makes no assumptions about how the effects in the different studies are related, it may not be used to draw any conclusions about a new study measuring the same outcome, much less about a new individual. Therefore, we will instead focus on the random effects model. This model allows the true means to differ between study populations but assumes they follow some underlying distribution, which is most often a normal distribution [69, 70]. It is common to interpret this assumption as meaning that each study in the meta-analysis was randomly sampled from a population of theoretically possible studies, including the population studied and methods of measurement. However, Higgins et al. [73] point out that this stronger assumption is often violated, as later studies are often designed based on the results of previous studies. However, Higgins et al. [73] focus on random effects models comparing treatment effects in two groups, while we focus on estimating normal ranges from a group of healthy subjects. Thus, this assumption may be reasonable.

3.2.2 Notation

Let $\bar{y}_i$ denote the observed mean for study $i = \{1, ..., N\}$, θ_i be study i's true mean, μ_{RE} be the overall mean of the distribution of study means, and σ_i^2 be study i's within-study variance. Also, let τ^2 be the variance of the θ_i across studies. Then, we have

$$\bar{y}_i \sim N(\theta_i, \sigma_i^2/n_i),\ \theta_i \sim N(\mu_{RE}, \tau^2) \tag{3.1}$$

In the frequentist framework, the overall mean μ_{RE} is traditionally estimated as a

weighted average of the study-specific means [69, 74]:

$$\hat{\mu}_{RE} = \frac{\sum_{i=1}^{k} y_i w_{i,RE}}{\sum_{i=1}^{k} w_{i,RE}}, \text{ for } w_{i,RE} = \frac{1}{s_i^2/n_i + \hat{\tau}^2}, \tag{3.2}$$

where s_i^2 is study i's within-study sample variance though there has been some debate about how to estimate τ^2. Here, we use the restricted maximum likelihood (REML) estimate, as implemented in the "meta" package [75]. The commonly-used estimate originally proposed by DerSimonian and Laird [74] has been shown to underestimate the true between-study variance, particularly when the number of studies k is small [76–78]. The overall variance in $\hat{\mu}_{RE}$ can be estimated by $V_{RE} = \frac{1}{\sum_{i=1}^{k} w_{i,RE}}$. The following is commonly used as an α-level confidence interval for $\hat{\mu}_{RE}$: $\hat{\mu}_{RE} \pm z_{\alpha/2} \times \sqrt{V_{RE}}$, where $z_{\alpha/2}$ is the standard normal critical value for the chosen significance level (α).

Alternatively, one can take a Bayesian approach and place prior distributions on μ_{RE} and τ as described in Section 3.3.2 [41, 73]. Because we consider the fixed mean assumption in the common effect model inappropriate in most situations, we use a random effects model for the two model-based approaches presented below. However, one could easily alter these methods to reflect a common effect assumption.

3.3 Methods for estimating the reference range from a meta-analysis

In estimating a 95% normal reference range, we aim to find an interval that contains approximately 95% of individuals in the target population [79, 80]. Because the models described in Section 3.2 only allow inference on the pooled mean, we need additional methods and assumptions to estimate the 95% normal reference range for an individual. First, we must make an assumption about the distribution of the data within each study, assuming we do not have access to the study's individual patient data (IPD). In this paper, we assume the individual-patient data in each study were generated from either a normal or log-normal distribution, with the family of distribution being consistent across all studies. We present

three approaches to estimating the normal reference range: a frequentist approach, a fully Bayesian approach, and an empirical approach. The frequentist and Bayesian methods assume the within-study distributions have the same variance in all studies, while the empirical approach does not. We present each of our three proposed methods under normality, and then show how to apply them under a log-normality assumption.

3.3.1 A frequentist approach

Under the random effects model, if we assume observations within each study are normally distributed, the within-study variances are the same in all studies, and the study-specific means follow a normal distribution, then we have $y_{ij} \sim N(\mu_{RE}, \sigma_T^2)$, where $\sigma_T^2 = \tau^2 + \sigma^2$ and σ^2 is the common within-study variance. We can estimate $\hat{\mu}_{RE}$ and $\hat{\tau}$ as described in Section 3.2, e.g., using REML, and estimate σ^2 as the unbiased pooled sample variance:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)}. \tag{3.3}$$

Substituting $\hat{\mu}_{RE}, \hat{\tau}^2$, and $\hat{\sigma}^2$ into the marginal distribution of y_{ij}, the marginal distribution of individuals, marginal to studies, can be estimated as $N(\hat{\mu}_{RE}, \hat{\sigma}_T^2)$, where $\hat{\sigma}_T^2 = \hat{\tau}^2 + \hat{\sigma}^2$. The $\alpha/2$ and $1-\alpha/2$ percentiles of this distribution can then be taken as the bounds of the α-level normal reference range: $\hat{\mu}_{RE} \pm z_{1-\alpha/2}\sqrt{\hat{\sigma}^2 + \hat{\tau}^2}$. Because $\hat{\sigma}^2$ and $\hat{\tau}^2$ were estimated, this suggests that a t-distribution may be appropriate instead of a normal. However, most meta-analyses will have a large enough total sample size across studies that the appropriate t-distribution will be closely approximated by a normal distribution. Alternatively, the fully Bayesian method presented in Section 3.3.2 accounts for the uncertainty in $\hat{\sigma}^2$ and $\hat{\tau}^2$.

3.3.2 A Bayesian approach

A fully Bayesian approach places prior distributions on μ_{RE} and τ. As in the frequentist approach, we assume that the true variances are the same in all studies, and now we use the

normal-theory sampling distribution of the sample variance to capture uncertainty about the within-study variance σ^2, according to this model:

$$\begin{aligned} \bar{y}_i &\sim N(\theta_i, \sigma^2/n_i) \\ \theta_i &\sim N(\mu_{RE}, \tau^2) \\ (n_i - 1)s_i^2 &\sim gamma\left(\frac{n_i - 1}{2}, \frac{1}{2\sigma^2}\right). \end{aligned} \tag{3.4}$$

We place a N(0,1000) prior on μ_{RE} and Unif(0,100) priors on τ and σ, then sample from the posterior predictive distribution for a new individual to incorporate into the normal reference range uncertainty about each of the parameter estimates:

$$y_{new} \sim N(\mu_{RE}, \sigma^2 + \tau^2), \tag{3.5}$$

where the predictive density of y_{new} given the data $\{y_{ij}\}$ is given by:

$$f(y_{new}|\{y_{ij}\}) = \int\int\int f(y_{new}|\mu_{RE}, \sigma^2, \tau^2) f(\mu_{RE}, \sigma^2, \tau^2|\{y_{ij}\}) d\mu_{RE} d\sigma^2 d\tau^2 \tag{3.6}$$

The limits of the α-level normal reference range can then be estimated by the $\alpha/2$ and $1 - \alpha/2$ percentiles of y_{new}'s predictive distribution.

3.3.3 An empirical approach

The third approach is a simple empirical approach that does not assume the studies all have the same within-study variances and does not specify the distribution of y_{ij} within each study. However, like the frequentist approach in Section 3.3.1, it does not account for estimation uncertainty and assumes the population captured across all studies follows a normal distribution. First, estimate the overall mean across all studies, weighted by study sample size:

$$\hat{\mu}_{emp} = \frac{\sum_{i=1}^{N} n_i \bar{y}_i}{\sum_{i=1}^{N} n_i}. \tag{3.7}$$

This is equivalent to the pooled mean in Laird and Mosteller's [64]'s fixed effects model, weighted by sample size. Then estimate the marginal variance across studies using the conditional variance formula $Var(Y) = E[Var(Y_{ij}|S = i)] + Var[E(Y_{ij}|S = i)]$:

$$\hat{\sigma}^2_{T,emp} = \frac{\sum_{i=1}^{N}(n_i - 1)s_i^2}{\sum_{i=1}^{N}(n_i - 1)} + \frac{\sum_{i=1}^{N}(n_i - 1)(\bar{y}_i - \hat{\mu})^2}{\sum_{i=1}^{N}(n_i - 1)} \tag{3.8}$$

The limits of the α-level normal reference range are then given by the $\alpha/2$ and $(1-\alpha/2)$ percentiles of a $N(\hat{\mu}_{emp}, \hat{\sigma}^2_{T,emp})$ distribution: $\hat{\mu}_{emp} \pm z_{1-\alpha/2} \times \hat{\sigma}_{T,emp}$. Conceição et al. [9] used this method but weighted by n rather than $n-1$ in the variance calculation. We prefer the unbiased estimate of the variance, but weighting by n will generally give similar results.

3.3.4 Lognormal distribution for y_{ij}

Each of the above methods can also be applied when each study's observations are assumed to be drawn from a $lognormal(\theta_i, \sigma_i^2)$ distribution, so that $log(y_{ij}) \sim N(\theta_i, \sigma_i^2)$. In this case, first transform the observed study means and sample variances to the log scale using Equation (3.9) before estimating the reference range:

$$\begin{aligned} \bar{y}_i^* &= \log\left(\frac{\bar{y}_i}{\sqrt{1 + \frac{\frac{n_i-1}{n_i}s_i^2}{\bar{y}_i^2}}}\right). \\ s_i^{2*} &= log\left(1 + \frac{\frac{n_i-1}{n_i}s_i^2}{\bar{y}_i^2}\right) \end{aligned} \tag{3.9}$$

This transformation uses the method of moments estimators for the location and scale parameters of the lognormal distribution. For more details, see the Appendix. The normal reference range can then be estimated as before, substituting $\bar{y}_i^*$ and s_i^* for the observed study-level means and standard deviations. Finally, exponentiate the limits of the resulting range to give the normal reference range: $\left(e^{\hat{\mu}-z_{1-\alpha/2}\hat{\sigma}_T}, e^{\hat{\mu}+z_{1-\alpha/2}\hat{\sigma}_T}\right)$. This method requires that the $\bar{y}_i^*$'s be normally distributed, an assumption that should be checked using a method

such as a Q-Q plot. Depending on the distribution of $\bar{y}_i$, the distribution of the $\bar{y}_i^*$'s can be quite skewed.

3.4 Simulations

3.4.1 Methods of simulation

To assess how well each of the three methods captures a true 95% normal reference range, we conducted simulations under a variety of different conditions. In all conditions, we assumed the true distributions within studies were normal and that the true study-specific means varied according to the random effects model in Section 3.2. For each condition, we then considered different values of true between-study variation (τ^2) as a proportion of total variability ($\tau^2 + \sigma^2$). In all conditions, each study had 50 subjects, with the total number of studies (N) being 5, 10, 20, or 30. The overall pooled mean (μ) was set to 8 and the true total variance ($\sigma^2 + \tau^2$) was 1.25 for all conditions. We conducted 1000 simulations for each condition. We considered scenarios where the true study-level variances were equal, as well as cases where they were not equal. For the frequentist approach, we used the R package "metagen" [81] to fit the random effects model using REML. For the Bayesian approach, we used JAGS version 4.3.0 with the packages "rjags" [51] and "coda" [52], in R version 3.6.0 [82]. We ran two chains each with 10,000 samples and after discarding 1,000 samples for burn-in.

Equal variances

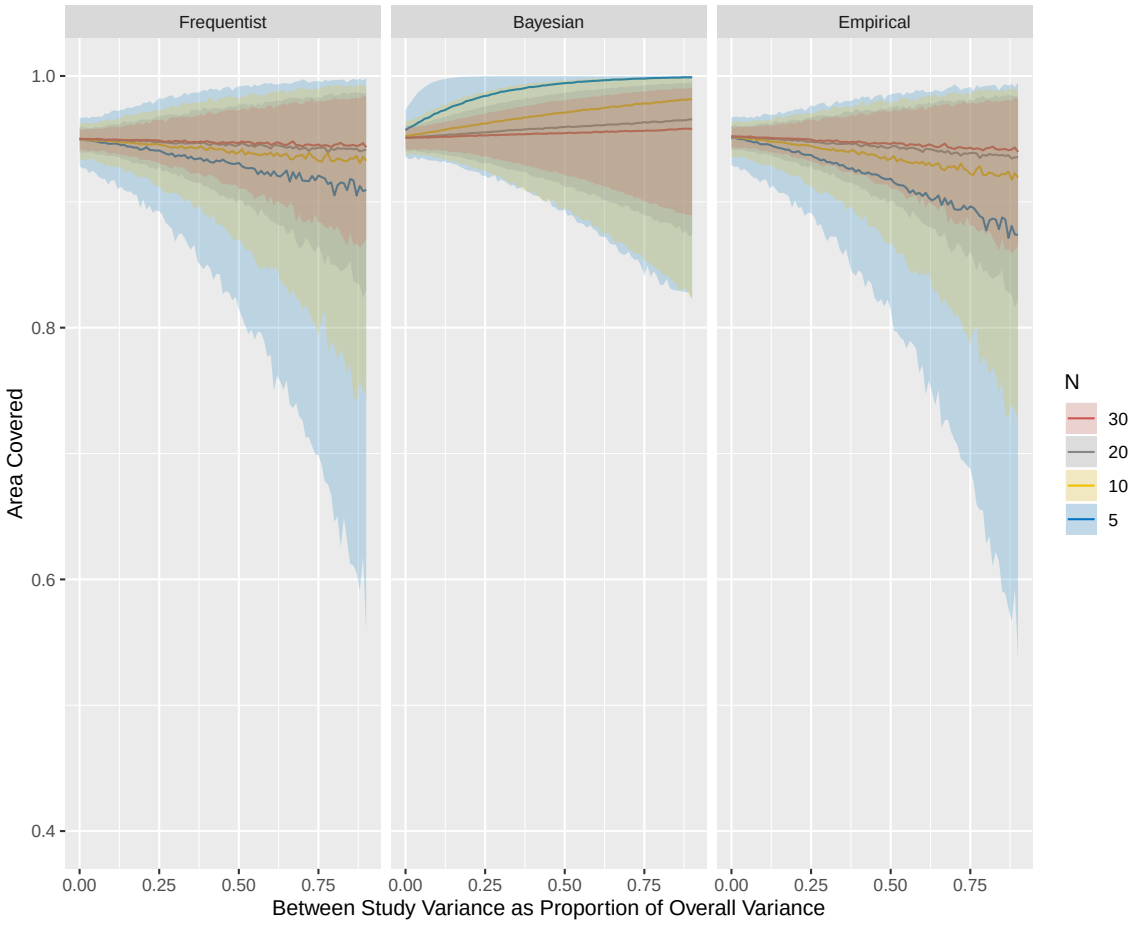
For the equal variance scenario, we first generated the true study-level means (θ_i's) according to a $N(\mu, \tau^2)$ distribution. For each study i, we then generated the individual-level data according to a $N(\theta_i, \sigma^2)$ distribution, where σ^2 was constant across studies. We summarized the means and standard deviations for each study to give the observed summary data, fit each of the three models, frequentist, Bayesian, and empirical, then found the area under

the probability density function of a $N(\mu, \sigma^2 + \tau^2)$ distribution between the upper and lower limits of the estimated 95% normal reference ranges.

Unequal variances

Data were simulated in the unequal variance scenario as in the equal variance scenario except that we generated σ_i, the true within-study standard deviation, from a doubly-truncated normal distribution, with both the left truncation point and mean equal to X and the right truncation point equal to X + 1, for X ranging from 0 to 0.64, with increments of 0.02. For each X, we estimated $E[\sigma_i^2]$ by simulating from the doubly-truncated normal distribution. These estimates ranged from 0.291 to 1.246. We let $\tau^2 = 1.25 - \hat{E}[\sigma_i^2]$ so as X increased, $\hat{E}[\sigma_i^2]$ increased as a proportion of the total variance. Because we truncated the normal distribution, the variance of σ_i remained constant throughout all conditions. We approximated the true reference distribution for y_{ij} by simulating from the full conditional distributions: $\sigma_i|X$, $\theta_i|\mu, \sigma_i$, and $y_{ij}|\theta_i, \sigma_i, \tau$.

Figure 3.1: **Simulation Results, Equal Variances.** Median, 2.5th percentile, and 97.5th percentile of the proportion of the true population distribution captured by the estimated 95% reference range, for different numbers N of studies. The horizontal axis is τ^2 as a proportion of the total variance.

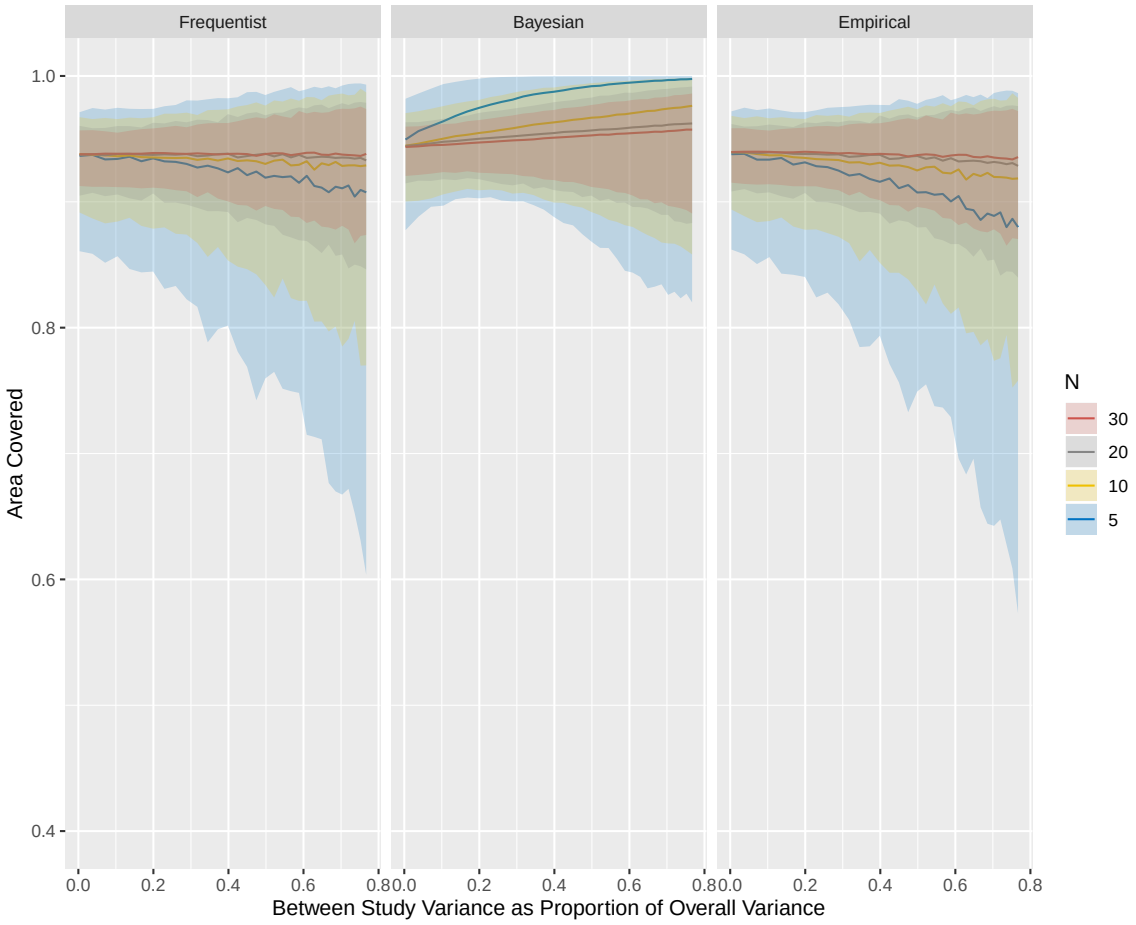


3.4.2 Simulation results

We first generated the data under the equal within-study variance scenario and measured the fraction of the true population distribution captured by each of the three reference

range methods, which we call the "coverage" (Figure 1). For example, when the between study variance comprised 25% of the overall variance, the frequentist reference ranges based on 1000 simulated meta-analyses containing 30 studies captured a median of 94.9% of a $N(\mu, \sigma^2+\tau^2)$ distribution, the true distribution of individual measurements. This can also be interpreted as the frequentist reference ranges excluding a median of 5.1% of extreme values. The observed median and variability of coverage depended on the true ratio of between-study variance (τ^2) to total variance ($\tau^2 + \sigma^2$) and the number of studies N included in the meta-analysis. For the frequentist and empirical methods, the median coverage decreased as τ^2 increased as a fraction of the total variance; this decrease was most pronounced when the number of studies included was small ($N = 5$ or 10). In these conditions, the empirical method's coverage decreased more quickly than the frequentist method's. Also, variation in coverage increased as τ^2 increased and decreased as N increased. While the variation in coverage also increased with τ^2 for the Bayesian posterior predictive interval, this effect was less dramatic. In contrast with the frequentist and empirical methods, the Bayesian method's median coverage increased with τ^2. This increase began for smaller τ^2 and was more extreme for small N. This increase in variation with τ^2 appears to reflect the additional estimation uncertainty when τ^2 is large, particularly when N is small. Unlike the frequentist and empirical methods, the Bayesian method accounts for posterior uncertainty about each parameter and thus appears more conservative. The results for the unequal within-study variances case were qualitatively similar to the equal variance case (Figure 3.2).

Figure 3.2: **Simulation Results, Unequal Variances.** Median, 2.5th percentile, and 97.5th percentile of the proportion of the true population distribution captured by the estimated 95% reference range, for different numbers N of studies. The horizontal axis is τ^2 as a proportion of the total variance.

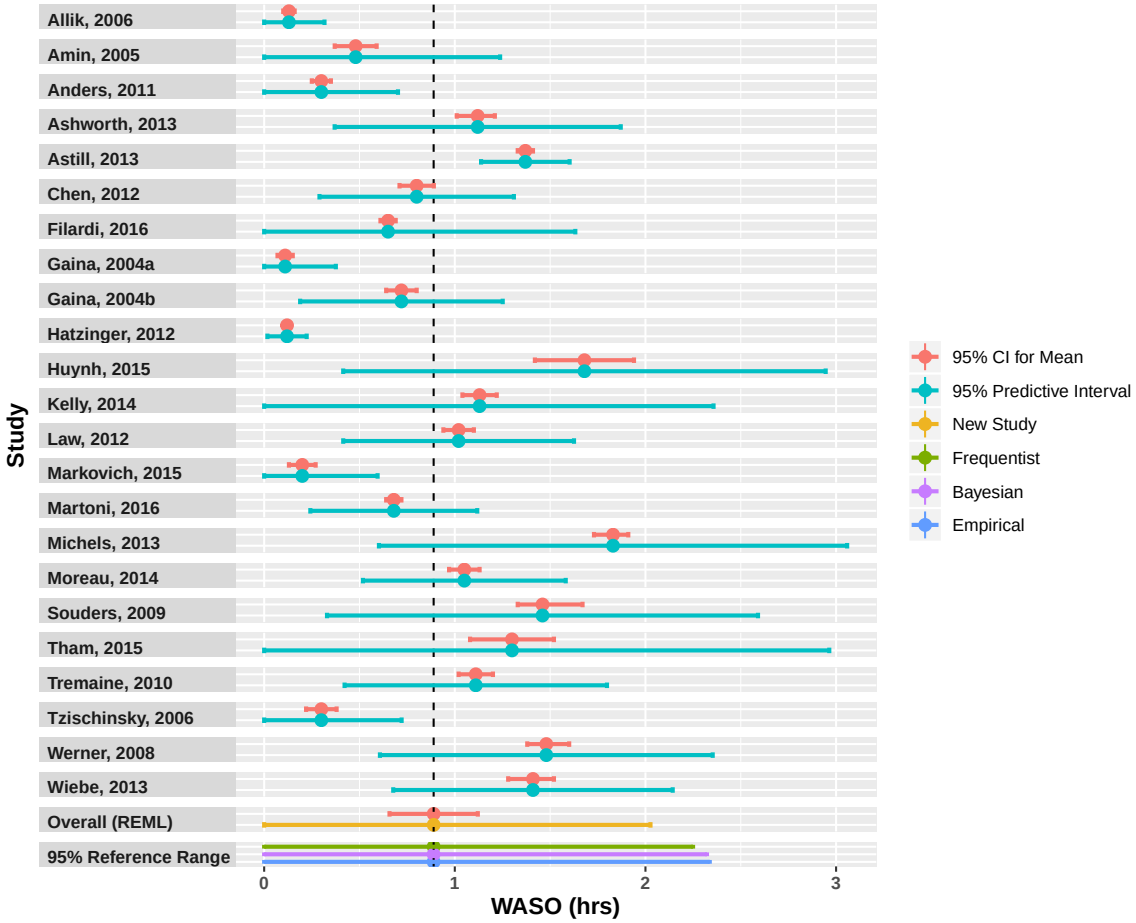


3.5 Examples

3.5.1 Example 1: Pediatric Nighttime Sleep

Galland et al. [12] sought to establish "reference values" for pediatric nighttime sleep outcomes measured by actigraphy, based on a systematic review and subsequent meta-analysis of 79 studies. We focus on the outcome wake after sleep onset (WASO) time in hours. The authors found 24 studies reporting WASO, with most participants belonging to the same age group (9-11 years) so they focused on the pooled mean across all age groups. In our review of these studies, one study included in this meta-analysis [83] did not appear to actually report WASO but rather reported the average length of wake bouts. We excluded this study from our analysis, which therefore contained 23 studies. Figure 3.3 shows the pooled mean and corresponding standard errors. In this case, only one of the 95% confidence intervals for the study means overlapped with the point estimate of the pooled mean. Galland et al. [12] explain that this variability reflects inconsistency across studies in how waking bouts were defined as well as a low specificity when using actigraphy to identify wakefulness. The authors also used meta-regression to investigate regional differences in sleep as a source of variation but did not observe a difference across study regions. This large variation in estimated WASO time across studies provides further evidence that the pooled mean may not provide a full picture of what constitutes a "normal" WASO time and that a reference range may be more useful. To better visualize the heterogeneity in WASO time within and across studies, we also present frequentist 95% prediction intervals based on a t-distribution for each study in the same figure [65, 66]. Because Galland et al. [12] only reported study means and standard errors, we obtained the standard deviations directly from each study's paper. When the paper did not report the standard deviation, we estimated the standard deviation using the standard error reported by Galland et al. [12] and a normal approximation. Therefore, our results should be interpreted as merely an illustration of the proposed methods.

Figure 3.3: **WASO** Mean (95% CI) and 95% predictive interval for a new individual for each study, overall estimate of pooled mean (95% CI) based on REML , 95% predictive interval for a new study mean, and 95% reference ranges based on Bayesian, empirical, and frequentist methods.



We checked whether the study means deviated from normality using a QQ-plot (see Supplementary Materials); no apparent departure from normality is observed except a few points at the end of both tails. As in the simulations, we used the R package "metagen" [81] to fit the random effects model using REML. We also used this to estimate the pooled mean

across studies and to obtain the prediction interval for a new study [73]. For the Bayesian approach, we again used JAGS version 4.3.0 with the packages "rjags" [51] and "coda" [52], in R version 3.6.0 [82]. We ran two chains each with 50,000 samples and after discarding 5,000 samples for burn-in. Convergence was assessed using MCMC standard error and visual inspection of trace plots.

The estimated 95% normal reference ranges were (-0.47, 2.24), (-0.54, 2.32), and (-0.33, 2.34) for the frequentist, Bayesian, and empirical methods, respectively. We truncated these at zero because negative WASO values are meaningless, giving (0, 2.25), (0, 2.32), and (0, 2.34). Based on the frequentist result, we would expect about 95% of healthy children to have WASO time between 0 and 2.25 hours based on actigraphy. This reflects the large amount of variability between individuals included in the meta-analysis. Before truncation, the Bayesian reference range was widest, followed by the frequentist reference range; the empirical method gave the narrowest interval. This is consistent with simulation results and is likely due to the Bayesian method accounting for uncertainty about the parameters σ, τ, and μ. The code and results for both case study examples are included in the Supplementary Materials.

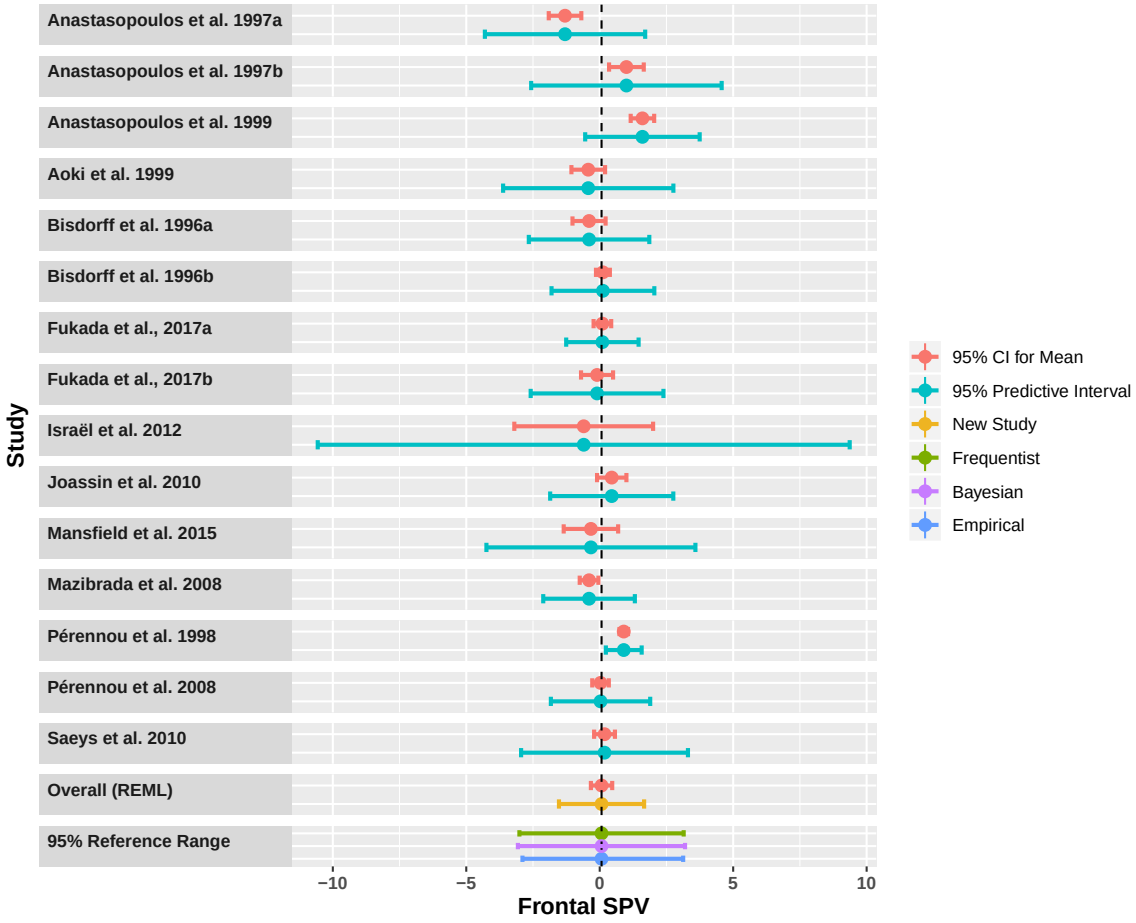
3.5.2 Example 2: Frontal SPV

Accurate perception of verticality is an important part of everyday functioning and can be altered in individuals such as "aged people, patients with vestibular disorders, Parkinson's disease, idiopathic scoliosis, and stroke patients" [9]. Accurate perception of verticality has also been associated with better functioning in patients following a stroke [84]. A person's subjective postural vertical (SPV) can be measured by placing them in a tilting chair while blindfolded and asking them to tell an examiner how to adjust the chair so they perceive that they are in an upright position. Frontal and sagittal SPV refer to deviation (in degrees) of the specified position from true verticality in the frontal and sagittal planes. Because SPV measurements can be used to assess neurological functioning, it is important to establish a

reference range in healthy persons.

In their meta-analysis, Conceição et al. [9] sought to establish reference ranges for frontal and sagittal SPV from 15 studies measuring frontal SPV and 5 studies measuring sagittal SPV. They estimated the reference range using the empirical approach except that they weighted by n rather than $n-1$ in the variance calculation. We re-analyze the data for frontal SPV using REML to estimate the pooled mean across all studies and the 95% predictive interval for a new study. We then used the same methods as in Section 3.5.1 to estimate the three reference ranges.

Figure 3.4: **Frontal SPV** Mean (95% CI) and 95% predictive interval for a new individual for each study, overall estimate of pooled mean (95% CI) based on REML , 95% predictive interval for a new study mean, and 95% reference ranges based on Bayesian, empirical, and frequentist methods.



We again checked for non-normality of the study-means using a Q-Q plot (see Supplementary Materials); no apparent departure from normality is observed except a few points at the upper tail. Figure 3.4 presents the reference range results as well as the estimated pooled mean and predictive interval for a new study. Conceição et al. [9] estimated the frontal SPV

reference range as $(-2.87°, 3.11°)$. The frequentist, Bayesian, and empirical methods gave estimated reference ranges $(-2.92°, 3.15°)$, $(-3.07°, 3.20°)$, and $(-2.89°, 3.13°)$, respectively. As expected, the empirical method's results are quite similar to those reported by Conceição et al. [9], while the frequentist and Bayesian methods give slightly wider intervals.

3.6 Discussion

This paper proposes three methods of estimating reference ranges for an individual from a meta-analysis. The methods are simple to implement and can serve as a starting point for future development. Based on the simulations, all three methods tended to perform best when the number of studies was large and between-study variability was relatively small. However, while the frequentist and empirical methods tended to underestimate the width of the reference range as between-study heterogeneity increased (particularly for small N), the Bayesian posterior predictive interval did not. This is likely because the Bayesian method accounts for estimation uncertainty about the parameters, while the other methods do not. Instead, the posterior predictive interval more often overestimated the width of the interval. Depending on how the reference range is used, one might consider this behavior conservative. We recommend using caution when the number of studies is small, such as 5 or 10. If the number of studies is very small and the estimated between-study variation makes up more than 50% of total estimated variation, it may be more useful to report reference ranges specific to each study, rather than a pooled range.

As for which method might be most appropriate in which circumstance, the simulation results suggest that when the number of studies is large (at least 20), and the normality assumptions hold, the three methods will likely perform similarly. Conceição et al. [9] used the empirical approach but weighted by n instead of $n - 1$. This suggests that the calculations required are intuitive and could easily be implemented by clinicians. The frequentist method, by contrast, requires familiarity with random effects models, although investigators are often interested in the pooled mean and thus likely to use this approach

anyway. Finally, the Bayesian predictive interval requires familiarity with Bayesian methods and a software such as JAGS, though it is still a simple model to implement and can account for estimation uncertainty.

Each method makes distributional assumptions beyond those needed when estimating the pooled mean. Besides the usual assumption made when using likelihood methods to analyze random effects models — that the study-means are normally distributed — the frequentist and Bayesian approaches also assume a normal distribution for individuals within a study. While this may appear problematic, prediction intervals for a new observation based on a single study regularly impose this assumption [85]. Unfortunately, if only study means and standard deviations are available, this assumption cannot be validated. Section 3.3.4 extended our approaches to allow individuals within studies to be lognormally distributed, but we caution that the transformed means on the log scale must be approximately normally distributed. This paper focuses on meta-analyses in which individual participant data (IPD) are not available; when IPD are available, non-parametric approaches using order statistics may be possible, as they are currently used in non-parametric estimates of reference ranges based on single studies [20, 21].

Another key assumption of the frequentist and Bayesian methods is that the true within-study variances are the same in all studies and that any observed differences are due to sampling variability. Differences between studies in sampling methods or measurement techniques could render this assumption invalid. However, Section 3.4.2's simulation results suggest that these methods may be robust to deviations from this assumption when the true study-specific standard deviations vary between studies according to a truncated normal distribution. Further work is needed to assess the models' performance under other deviations from this assumption.

Finally, we reiterate that random effects models for estimating the pooled mean, on which we built the frequentist and Bayesian methods, require the study-specific means to be normally distributed. This is true of most random effects methods, except for the

method of moments estimator developed by DerSimonian and Laird [74], which is known to underestimate between-study variability and therefore give results with inappropriately high precision [76–78]. It is a common misconception that normality of the study means is guaranteed by the central limit theorem (CLT) [68]. At best, the CLT only ensures that the sampling distribution of an observed average for a single study has an approximate normal distribution, not that the true means of the collected studies follow a normal distribution. One way of assessing departures from this assumption is the use of a normal Q-Q plot.

Although our methods make specific distributional assumptions, they do provide a starting point for additional development. Future work should generalize these methods to address instances where the assumptions used here are likely not met. This could involve cases with or without IPD. Section 3.4.2's simulation results show that our approaches work well in cases with many studies and relatively low between-study variability. However, future studies should compare these methods with new methods incorporating IPD. Future methods should also improve performance when the number of studies is small or the between-study variation is large. The proposed methods may also be extended to cases where the data from each study are assumed to follow truncated normal distributions. Finally, these methods could be extended to a meta-regression setting to include characteristics such as age or sex, at either the individual or the study level.

Chapter 4

A Guide to Estimating the Reference Range from a Meta-Analysis Using Aggregate or Individual Participant Data

4.1 Clinical Scenario

A 50-year-old healthy man without significant past medical history presents to his primary care physician for a preventive health exam. He is concerned because his sister was diagnosed with "liver fibrosis". Although the patient has normal liver transaminases, INR, platelets, and albumin levels, he has been overweight his entire life and drank alcohol heavily during his college days. A liver biopsy, which is the gold standard diagnostic tool, is too invasive and costly to be performed on a healthy asymptomatic individual. A noninvasive ultrasound-based test called transient elastography was introduced in 2003. However, the normal range for this test is not known and has been reported from several heterogeneous studies of

patients with various races, ethnicities, and other demographic characteristics. Bazerbachi et al. [7] conducted a systematic review and meta-analysis of these studies measuring liver stiffness in healthy adults where individual participant data were available. The authors estimated the mean stiffness in healthy non-obese individuals and reported the confidence interval for the mean as the reference range, despite this interval reflecting only uncertainty in the pooled mean rather than the variation across individuals. We revisit their analysis in order to construct a reference range that incorporates natural variability across healthy individuals in addition to the uncertainty in the estimated mean.

4.2 Introduction

Often clinicians would like to know whether a patient's measurement falls within some "normal" range for healthy individuals. While meta-analysis most frequently involves summarizing one or more treatment effects on an outcome, there are many examples of meta-analyses of normative data [7–9,12,54,56,58,60–63,86,87]. Normative data are assumed to be drawn from a predefined healthy population (e.g., with certain inclusion and exclusion criteria) that can serve as a reference for future comparison [15]. Therefore, these studies aim to establish "normal" values for continuous measurements using data from healthy individuals across multiple studies. These data may be drawn from normative studies of healthy individuals, cohort studies, the control arms of case-control studies, or baseline values from randomized-controlled trials in healthy populations [7, 9, 12]. In most cases, a reference range, or an interval in which we would expect the measurements of a specified proportion of a healthy population (e.g., 95%) to fall [20, 21], would provide the most information in determining whether a patient's measurement is "normal." This can also be defined as a prediction interval for the value of a new healthy individual conditional on the normative data from existing evidence [21]. While several studies in the biomedical literature have used ad-hoc methods to report reference ranges estimated from meta-analyses consistent with this definition [9, 56, 58, 60], we have recently proposed three methods for estimating

the reference range from a meta-analysis with aggregated data [88]. Here to provide some practical guidance, we describe how to calculate the reference range from a meta-analysis and outline how it differs from the confidence interval for the pooled mean and the prediction interval for the mean of a new study [67, 73]. We provide an overview of the three methods and apply them to a systematic review and meta-analysis of studies measuring normative liver stiffness in adults. We consider using aggregate data from publications, but also extend this to using individual participant data.

4.3 What aggregate data are typically available and needed for a reference range meta-analysis?

Often, when conducting a meta-analysis of multiple studies to estimate the reference range, only aggregate data are available from published studies. The required aggregate data typically include the observed means, standard deviations, and sample sizes from each study. Studies may also report demographic information, such as the proportion of males and females, or the mean age of participants in the study.

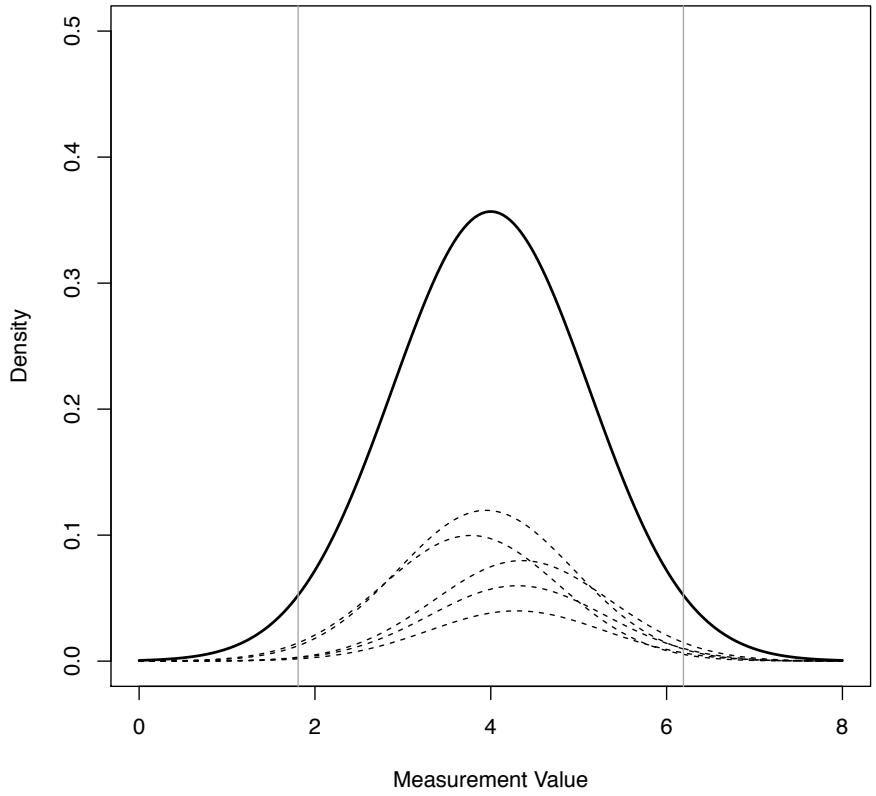
4.4 Defining the population of interest

To determine whether the studies included in a meta-analysis have enrolled participants who belong to the pre-specified target population for which a reference range is being sought, we suggest evaluating two sources of information. The first is the inclusion and exclusion criteria of the meta-analysis. The second is the observed demographic information provided in the manuscripts of included studies (e.g., mean age, proportion of males or females). Based on these two sources of data a judgment needs to be made about whether the studies include representative participants from the target population whose reference range is being sought. It is also important to consider whether some studies have enrolled participants with occult disease and exclude such studies. For example, healthy volunteers who have occult

fatty liver disease and enroll in hepatology studies is a well recognized phenomenon [89]. Thus, included studies should perform sufficient testing to rule out occult disease when possible.

Each of the proposed methods for estimating the reference range allow the underlying means of each study included in the meta-analysis to differ (this is also known as a "random effects" assumption). In other words, variation in the observed means across studies can be attributed to actual differences and sampling variability [67]. One can often achieve this by carefully defining certain inclusion and exclusion criteria in the systematic review. In particular, we assume that the studies included in the meta-analysis are a representative or random sample from a greater "superpopulation" of potential studies and are interested in the marginal (overall) distribution of individuals across all of these potential studies (Figure 4.1). Determining whether this sample is representative also requires investigating possible heterogeneity sources, as described in the next section. We prefer to focus on the overall distribution, rather than conditioning on a specific study, since it may be unclear which theoretical study population a patient who presents to a clinic would belong to in practice. These study-specific underlying populations would likely differ in size, though knowledge of these true population sizes is not necessary under the random effects assumption.

Figure 4.1: **Target Population** Marginal (overall) distribution and a selection of possible transient elastography liver stiffness measurement study populations according to a random-effects model where $\mu = 4, \sigma = 1, \tau = 0.5$. The distributions of study means and individuals within each study are all normal. Each of the meta-analysis methods presented allows for true differences between sub-populations, and the target population is the overall distribution that captures each of these.



In the clinical scenario described previously, the overall target population consists of healthy non-obese individuals without evidence of liver steatosis or fibrosis across all potential studies, as we aim to characterize liver stiffness measurements that would be extreme for patients with healthy livers while incorporating the full variability found across different populations of healthy patients. Therefore, studies of obese patients would have been

excluded when estimating the reference range.

4.5 Investigating sources of heterogeneity

The random effects assumption described earlier to account for between-study heterogeneity assumes that there are many possible studies and that the underlying study means follow a distribution, typically a normal distribution. This assumption is consistent with small variations across studies such as those due to slightly different but overlapping study populations, similar but not identical equipment, or different personnel collecting measurements. If it is believed that the overall population can be partitioned into several distinct subpopulations with different measurements, the random effects model would likely be inappropriate. Instead, separate reference ranges corresponding to each population would be more informative. For example, if measurements are suspected to vary by subgroups, such as biologic sex or age, separate reference ranges specific to these groups, based on stratified analysis or meta-regression, may be more clinically meaningful. Hypothesized sources of heterogeneity could be investigated using subgroup analyses or meta-regression methods [90], although this often lacks power. Because the overall mean and variance across individual participants are of equal interest when estimating the reference range, heterogeneity in the within-study variances should also be carefully explored. For example, differences in the variances in individual measurements within each study can be investigated visually using a forest plot of the observed study standard deviations and their corresponding confidence intervals.

4.6 Meta-analysis methods for estimating the reference range

We previously proposed three methods for estimating the reference range using aggregate data [88]; these methods are summarized in Table 4.1 and described in further detail in Appendix C. The first two methods, the frequentist method, and the Bayesian posterior

predictive interval, assume that 1) values of the variable of interest follow a normal distribution for each study population; 2) the variances of individual measurements within each study are equal across studies; 3) that the true study means are also normally distributed. These assumptions then imply that the overall distribution across studies is also normal.

The frequentist approach involves estimating the shared within-study variance, fitting a random-effects model on the aggregate data, and then using the estimated pooled mean and within and between-study variances to approximate the overall distribution of individuals. The bounds of the estimated 95% reference range are then given by finding the 2.5th and 97.5th quantiles of this overall normal distribution, assuming the estimated parameters are fixed quantities (i.e., ignoring their uncertainty).

The Bayesian method requires fitting a random effects model on the aggregate data where the shared within-study variance is estimated using the sampling distribution of the sample variance. The bounds of the 95% reference range are then given by the 2.5th and 97.5th quantiles of the posterior predictive distribution for a new individual. This differs from the other two methods in that the reference range becomes wider with greater uncertainty by considering the variation of parameters, consistent with the definition of the reference range as a prediction interval. While it may be possible to introduce this behavior with the frequentist approach using a t-distribution, the appropriate degrees of freedom are unclear and likely require approximation. Furthermore, the degrees of freedom will depend on both the estimated within and between-study variances and will likely be high when the number of studies is large or when the between-study variance is small relative to the total variance. Under those conditions, the t-distribution will strongly resemble that of a normal distribution, meaning that incorporating estimation uncertainty will make little difference in the width of the reference range. Additionally, depending on the application, it may be more prudent to flag a truly healthy individual as abnormal, thus necessitating more investigations, rather than failing to discern pathology in a sick patient. In such scenario, it may be preferable to omit the estimation uncertainty of parameters from the width of the

interval, because under the Bayesian approach, the estimated interval may contain greater than 95% of measurements in the case of large estimation uncertainty (e.g., when the number of studies is small, the between-study variance may be estimated with greater uncertainty). Conversely, if avoiding over-diagnosis is of greater concern, the estimated interval from the Bayesian approach may be preferred.

The frequentist and Bayesian methods also make the usual random effects assumption that the study means (random effects) follow a normal distribution [73]. It is often incorrectly assumed that the central limit theorem (CLT) guarantees this [68]. The CLT only guarantees normality of the sampling distribution of the mean from a single study, not the overall collection of study means. Instead, this assumption should also be visually assessed. Methods have also been developed for estimating prediction intervals for a new study effect that do not require this normality assumption, such as those based on bootstrap sampling methods [68, 91]. Future work could expand these methods to prediction on the individual level.

The third aggregate data approach, the empirical approach, does not require the data within each study to be normally distributed or equal within-study variances, only that the overall distribution across all studies is normal. Instead, the pooled mean is estimated as a weighted average of the study means, and the total variance is estimated as the sum of a weighted average of the sample variances, and the sample variance of the study means. This empirical method could also likely be used when the overall distribution is assumed to be any other distribution that is entirely determined by its mean and variance. Furthermore, while the methods mentioned thus far assume that the data within each study are normally distributed, we also describe in the Appendix how to handle aggregate data that are believed to follow a lognormal distribution.

Table 4.1: **Methods for Estimating the Reference Range**

1A. Frequentist Approach (Aggregate Data):

1) Estimate the pooled mean (μ_{RE}) and between-study variation (τ^2) using a frequentist random-effects model such as REML
2) $\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(n_i-1)s_i^2}{\sum_{i=1}^{N}(n_i-1)}$, where s_i^2 is the sample variance from study $i \in \{1, \ldots, N\}$
3) Limits of the estimated reference range: $\alpha/2$ and 1 - $\alpha/2$ quantiles of a $N(\hat{\mu}_{RE}, \hat{\sigma}^2 + \hat{\tau}^2)$ distribution

1B. Frequentist Approach (Individual Participant Data):

1) Fit a frequentist random-effects model (linear mixed model) directly using the individual participant data
2) $\hat{\tau}^2$ = Estimated variance of the random effects
3) $\hat{\sigma}^2$ = Estimated residual variance
4) $\hat{\mu}_{RE}$ = Estimated pooled mean (fixed effect)
5) Limits of the estimated reference range: $\alpha/2$ and 1 - $\alpha/2$ quantiles of a $N(\hat{\mu}_{RE}, \hat{\sigma}^2 + \hat{\tau}^2)$ distribution

1C. Bayesian Approach (Aggregate Data):

1) $\bar{y}_i \sim N\left(\theta_i, \frac{\sigma^2}{n_i}\right), \theta_i \sim N(\mu_{RE}, \tau^2), (n_i - 1)s_i^2 \sim gamma\ (\frac{n_i-1}{2}, \frac{1}{2\sigma^2})$
2) Place N(0, 1000) prior on μ_{RE} and Uniform(0, 100) priors on σ and τ
3) Use MCMC sampler (such as JAGS, Stan, or WinBugs) to sample from posterior predictive distribution for a new individual: $y_{new} \sim N(\hat{\mu}_{RE}, \hat{\sigma}^2 + \hat{\tau}^2)$
4) Limits of the estimated reference range: $\alpha/2$ and 1 - $\alpha/2$ quantiles of y_{new} samples

1D. Bayesian Approach (Individual Participant Data):

1) $y_{ij} \sim N(\theta_i, \sigma^2), \theta_i \sim N(\mu_{RE}, \tau^2)$
2) Place N(0, 1000) prior on μ_{RE} and Uniform(0, 100) priors on σ and τ
3) Use MCMC sampler (such as JAGS, Stan, or WinBugs) to sample from posterior predictive distribution for a new individual: $y_{new} \sim N(\hat{\mu}_{RE}, \hat{\sigma}^2 + \hat{\tau}^2)$
4) Limits of the estimated reference range: $\alpha/2$ and 1 - $\alpha/2$ quantiles of y_{new} samples

1E. Empirical Approach (Aggregate Data):

1) Empirically estimate the pooled mean and total variance:
$\hat{\mu}_{emp} = \frac{\sum_{i=1}^{N} n_i \bar{y}_i}{\sum_{i=1}^{N} n_i}$, $\hat{\sigma}^2_{T,emp} = \frac{\sum_{i=1}^{N}(n_i-1)\, s_i^2}{\sum_{i=1}^{N}(n_i-1)} + \frac{\sum_{i=1}^{N}(n_i-1)\,(\bar{y}_i-\hat{\mu}_{emp})^2}{\sum_{i=1}^{N}(n_i-1)}$
2) Limits of the estimated reference range: $\alpha/2$ and 1 - $\alpha/2$ quantiles of a $N\left(\hat{\mu}_{emp},\ \hat{\sigma}^2_{T,emp}\right)$ distribution

1F. Empirical Approach (Individual Participant Data):

1) Empirically estimate the pooled mean and total variance as the observed mean and variance of the pooled individual participant data:
$\hat{\mu}_{emp} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{N} n_i}$, $\hat{\sigma}^2_{T,emp} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i}(y_{ij}-\hat{\mu}_{emp})^2}{(\sum_{i=1}^{N} n_i)-1}$
2) Limits of the estimated reference range: $\alpha/2$ and 1 - $\alpha/2$ quantiles of a $N\left(\hat{\mu}_{emp},\ \hat{\sigma}^2_{T,emp}\right)$ distribution

Simulation results suggest that each of the proposed aggregate data approaches perform similarly when the between-study heterogeneity is relatively small, and the number of studies in the meta-analysis is large (at least 20) [88]. However, some caution should be used in cases of large between-study heterogeneity or very few studies. It may be more appropriate

to construct reference ranges separately for subgroups or studies of different populations in these situations. In particular, if there is unexplained between-study heterogeneity that comprises approximately 30-50% or more of the total estimated variance, it is important to consider the interpretability of the estimated reference range carefully. While the equal within-study variation assumption made by the frequentist and Bayesian methods is arguably quite strong, Siegel et al. demonstrated through simulations that these methods might be robust to small differences in the true variances across studies [88]. However, if the within-study variances plausibly differ according to some characteristic of the studies, such as the proportion of males vs. females, separate reference ranges for these groups may be more clinically meaningful regardless of the distributional assumptions of the method used.

4.7 Applied example

We now re-analyze the data used in the clinical scenario [7] in order to construct a reference range that reflects natural variability across healthy individuals rather than the uncertainty in the estimated pooled mean.

4.7.1 Defining the population of interest

Individuals were included in the original analysis if they had a BMI less than 30, and did not have hypertension, dyslipidemia, hepatic steatosis on ultrasound, or diabetes mellitus. This resulted in 3652 individuals across 20 studies. Because one of these studies only contained four individuals meeting the inclusion criteria, we further excluded these four patients from the analysis. This resulted in a final dataset containing 3648 individuals across 19 studies.

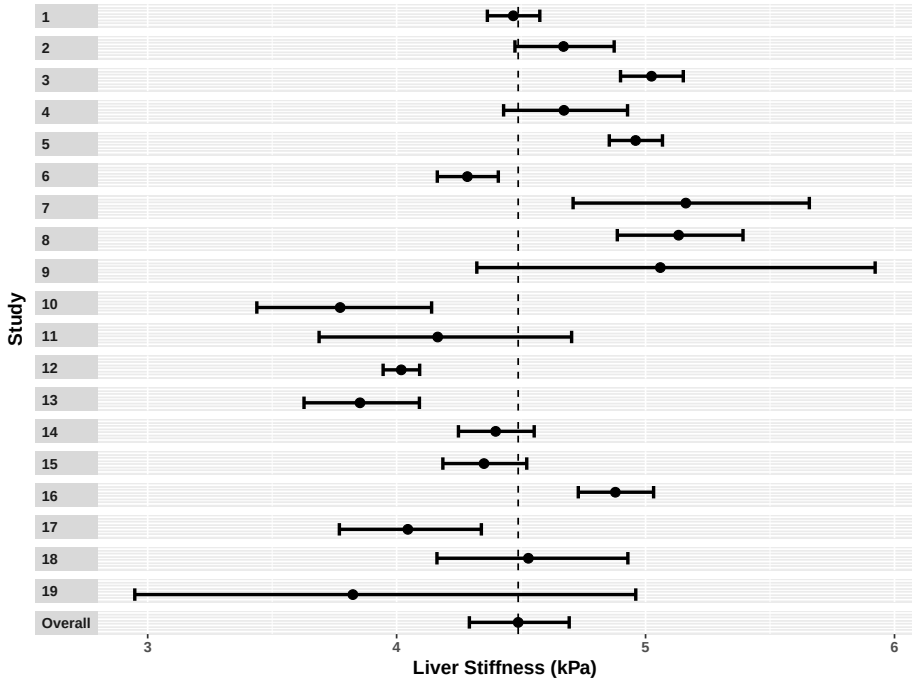
4.7.2 Derivation of aggregate data

In order to replicate the scenario where only aggregate data were available, rather than IPD, we summarized the data within each study by the mean, standard deviation, and sample size (Table C.2).

4.7.3 Application of methods

We present a typical forest plot for the study-specific means and pooled mean, with their corresponding confidence intervals in Figure 4.2. The pooled mean was estimated using the aggregate data and a frequentist random effects model (using REML estimation) implemented in the R package "meta" [75]. Because liver stiffness measurements cannot be negative and the observed distribution of measurements was slightly right-skewed, we first log-transformed the liver stiffness measurements and then exponentiated the results for the means and 95% confidence intervals. Because we were using aggregate data, this log-transformation required using the approximation described in the Appendix and Table C.1.

Figure 4.2: **Forest Plot of Study Means for Clinical Scenario** Estimated mean (95% confidence interval) for each transient elastography liver stiffness measurement study and estimated pooled mean (95% confidence interval) based on aggregate data. All calculations were completed on the log-scale, and the resulting estimates were exponentiated.



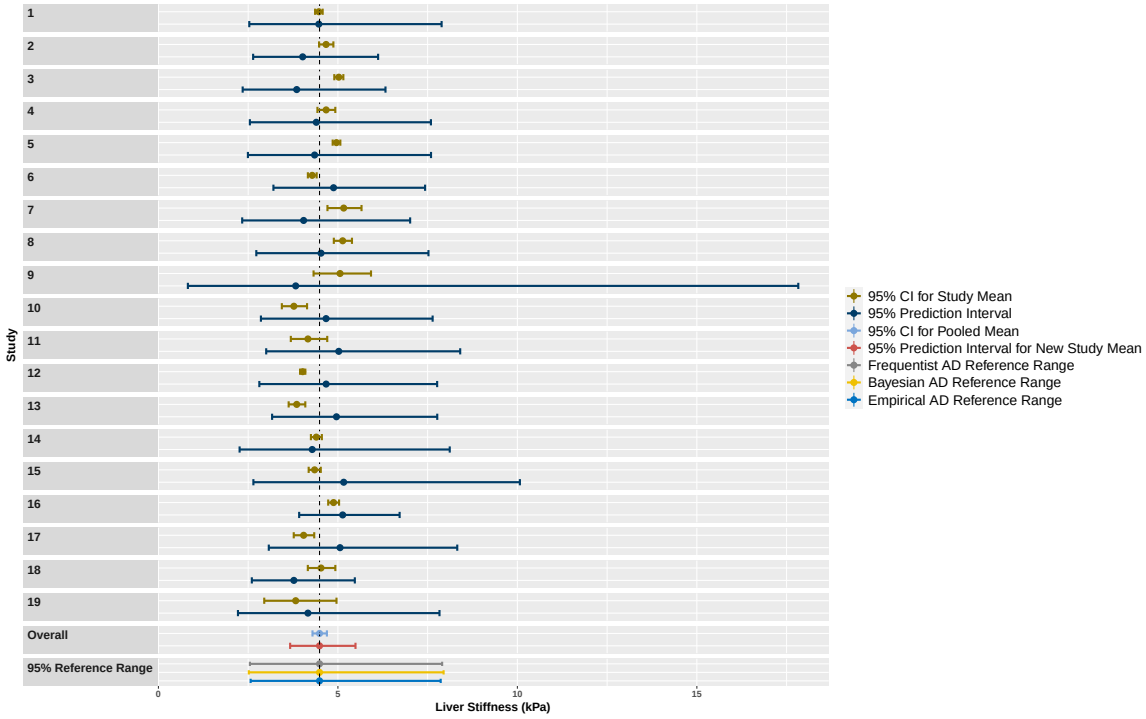
We next applied each of the proposed methods for estimating the 95% reference range (Table 4.2) using the aggregate data. As previously mentioned, we used the R package "meta" [75] to fit the frequentist random-effects model with aggregate data. We implemented the Bayesian models in JAGS using the R packages "rjags" and "coda" [51, 52]. For the Bayesian models, we ran two chains with 100,000 iterations each and a burn-in period of 5,000 iterations and assessed convergence based on trace plots, the MCMC error, and the potential scale reduction factor. All analyses were conducted using R version 3.6.3 [50].

Table 4.2: **Reference Range Results for Clinical Scenario with Aggregate Data** Estimated 95% reference ranges for liver stiffness measurement using each of the methods presented with aggregate data. The reference ranges were estimated on the log-scale, and the resulting intervals were exponentiated.

Method	**Estimated 95% Reference Range**
Frequentist	(2.55 kPa, 7.90 kPa)
Bayesian	(2.52 kPa, 7.94 kPa)
Empirical	(2.57 kPa, 7.86 kPa)

The estimated reference ranges were similar across each of the methods used (Table 4.2, Figure 4.3). The Bayesian posterior predictive interval was slightly wider, followed by the frequentist method, then the empirical approach. We would expect the Bayesian method to give a wider reference range as it incorporates uncertainty in the parameter estimates. Figure C.1, included in the Appendix, displays the observed standard deviations of the log of liver stiffness within each study and their respective 95% confidence intervals. This allows us to assess the equal within-study variance assumption imposed by the frequentist and Bayesian methods for estimating the reference range. In general, this figure shows that most of the observed study standard deviations are very similar and that there is a high degree of overlap in their respective confidence intervals. However, studies 9 and 16 look to have a slightly different standard deviations from the other studies. We therefore performed a sensitivity analysis where we removed studies 9 and 16 and compared the estimated reference ranges with and without them. The results of this sensitivity analysis are given in Table C.3 in the Appendix; the results with and without these two studies are similar.

Figure 4.3: **Forest Plot of Results for Clinical Scenario** 95% confidence interval for each study mean, 95% frequentist prediction interval for a new individual's transient elastography liver stiffness measurement by study, 95% confidence interval for the pooled mean, 95% prediction interval for a new study mean, and estimated 95% reference ranges using the four methods presented. All calculations were completed on the log-scale and the resulting estimates were exponentiated



Each of the estimated reference ranges can be interpreted as the predicted interval in which we would expect 95% of liver stiffness measurements of healthy individuals to fall. For example, based on the Bayesian individual participant data reference range, we would expect 95% of healthy patients to have liver stiffness measurements between 2.52 kPa and 7.94 kPa. Therefore, if our hypothetical patient who is concerned about his family history of liver fibrosis, had a liver stiffness measurement of 9.00 kPa, this may necessitate further

investigations, as this degree of liver stiffness is atypical of a healthy individual.

The 95% confidence interval for the pooled mean ([4.29 kPa, 4.69 kPa]), is much narrower than any of the estimated reference ranges. This demonstrates the difference that incorporating natural within-person variability makes when constructing the reference range. This is also true when comparing the estimated reference ranges to the frequentist 95% prediction interval for the mean of a new study: (3.67 kPa, 5.49 kPa) instead of the measurement of an individual [67,68]. This interval is wider than the 95% confidence interval for the pooled mean, as it reflects between-study variance and its corresponding estimation uncertainty. However, unlike the reference ranges, it still does not reflect within-study variation across healthy individuals. We can also compare the results to the 2.5th and 97.5th quantiles of the individual measurements, ignoring study assignment: (2.70 kPa, 7.49 kPa). The estimated reference ranges that incorporate study assignment are slightly wider than this because they allow for between-study variation and the possibility of more extreme measurements in a future study. The confidence interval for the pooled mean and the prediction interval for a new study mean are far narrower and do not capture healthy individuals' full variation. The different interpretations of the reference ranges, the confidence interval for the pooled mean, and the prediction interval for a new study are summarized in Table 4.3.

Table 4.3: **Comparison of Interpretations of Intervals Described in Paper**

Interval	Pooled Mean	95% Prediction Interval for a New Study	95% Reference Range (proposed)
Interpretation	**Frequentist 95% Confidence Interval (a, b):** "We are 95% confident that the mean across all studies is between a and b." **Bayesian 95% Credible Interval (a,b):** "The true mean across all studies lies between a and b with 95% probability."	**Prediction interval (c,d):** "The **mean of a new study** (from the same overall target population) is expected to fall between c and d with 95% probability."	**Reference range (e,f):** "The **measurement of a new individual** is expected to fall between e and f with 95% probability."
Assumptions	Under a random-effects model: • Study means follow a normal distribution[1] • The means are exchangeable across studies • The studies included are representative of some superpopulation of interest	Under a random-effects model: • Study means follow a normal distribution • The means are exchangeable across studies • The studies included are representative of some superpopulation of interest	**Frequentist:** • Measurements within each study follow a normal distribution • Study means follow a normal distribution and are exchangeable • Constant within-study variance **Bayesian:** • Same as frequentist **Empirical:** • Measurements across all studies follow a normal distribution
Estimation	**Frequentist:** $\hat{\mu}_{RE} \pm t_{N-1;1-0.05/2}\, SE(\hat{\mu}_{RE})$ (where N = # of studies) **Bayesian:** 2.5th and 97.5th quantiles of the posterior distribution of the pooled mean (μ_{RE})	**Frequentist:** $\hat{\mu}_{RE} \pm t_{N-2;1-0.05/2}\sqrt{\widehat{Var}(\hat{\mu}_{RE}) + \hat{\tau}^2}$ [67] **Bayesian:** 2.5th and 97.5th quantiles of the posterior predictive distribution of a new study: $N(\mu_{RE}, \tau^2)$, where μ_{RE} and τ^2 refer to their posterior distributions	**See Box 1**

[1] * The method proposed by DerSimonian and Laird [74] does not require the study means to be normally distributed but can underestimate the between-study variance, particularly when the number of studies is small [76]

4.8 Estimating the reference range using individual participant data (IPD)

All three approaches are designed for the meta-analysis of aggregate data, where only the study means, standard deviations, and sample sizes are known. Because of this, we also include how the reference range could be calculated using IPD without first aggregating the data (i.e., a one-step approach) (Table 4.1). These approaches are one-step analogs of each of the three approaches described previously, though the two versions of the empirical approach are in theory equivalent. The estimated reference ranges based on individual participant data ultimately serve as a "gold-standard".

Furthermore, individual participant data allows for a more detailed exploration of the modeling assumptions. Each of the methods previously discussed assumes that the individuals across all studies follow an overall normal distribution. Both the Bayesian and frequentist approaches also assume that the data within each study are normally distributed and that the within-study variances are equal across studies. Unlike with aggregate data, if IPD are available, these normality assumptions can be visually assessed using methods such as histograms and normal Q-Q plots. Because of this, access to IPD even for 1 or 2 studies could be valuable in investigating these distributional assumptions before using an aggregate data method to estimate the reference range. Similarly, with aggregate data, we cannot directly log-transform the individual measurements. Instead, the approximation given in the Appendix Table C.1 must be used.

4.8.1 Applied example with individual participant data

Here, we present the results for the clinical scenario using both the aggregate (two-step) approaches as well as the one-step frequentist and Bayesian approaches based on the IPD. In all cases, the data are first log-transformed (before aggregating) and the resulting ranges are exponentiated. While the aggregate data approaches are still valid even when individual participant data are available, the one-step (IPD) approaches are the gold standard.

Table 4.4: **Reference Range Results for Clinical Scenario with IPD** Estimated 95% reference ranges for liver stiffness measurement using IPD. The reference ranges were estimated on the log-scale and the resulting intervals were exponentiated

Method	Estimated 95% Reference Range
Frequentist AD	(2.62 kPa, 7.74 kPa)
Bayesian AD	(2.61 kPa, 7.79 kPa)
Empirical AD	(2.64 kPa, 7.69 kPa)
Frequentist IPD	(2.63 kPa, 7.72 kPa)
Bayesian IPD	(2.52 kPa, 7.94 kPa)
Empirical IPD	(2.64 kPa, 7.69 kPa)

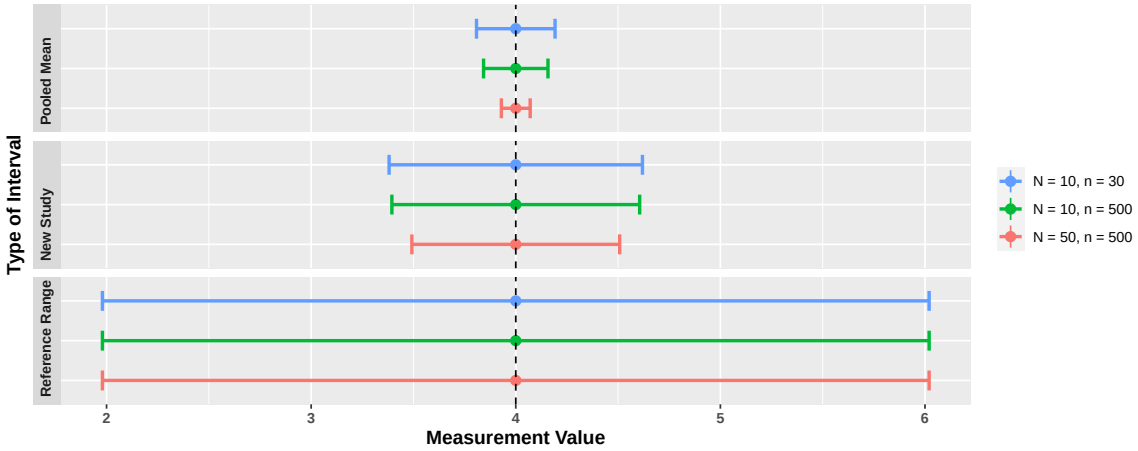
A histogram of the pooled log liver stiffness measurements (Figure C.3) as well as histograms by study (Figure C.4) are included in the Appendix and demonstrate no clear violations to the overall and within-study normality assumptions described previously. We were able to assess these normality assumptions using IPD, whereas this would not be possible with only aggregate data. As expected, the frequentist method using IPD gave a slightly narrower estimated reference range than the Bayesian method with IPD. With IPD available, we directly obtained the mean and standard deviation on the log scale for each study. This is different from computing the mean and standard deviation in the log scale using the methods presented in Appendix Table C.1 with only the reported mean and standard deviation in the original scale. Because the log-transformation differed between this analysis and the aggregate data analysis presented previously in Table 4.2, we would expect the results to be slightly different from the previous analysis even amongst the aggregate data approaches. Notably, the results using the aggregate data are comparable to those based on the IPD (Table 4.4). This supports the validity of the aggregate data approaches in this case, an important point given that IPD are rarely available for all studies included in a meta-analysis. We also repeated the sensitivity analysis described in the previous section

with the IPD, and again observed similar results with and without studies 9 and 16, as shown in the Appendix (Table C.4).

4.9 Interpretation of results

Because there has been little guidance in the literature on estimating reference ranges from a meta-analysis, many meta-analytical studies have reported the pooled mean as a "reference value" [8, 12, 53]. While the pooled mean can establish a point of reference, it does not capture natural variation across healthy individuals. As a result, some studies have also reported the 95% confidence interval for the pooled mean as a "reference range" [7, 61, 62], although this better reflects the uncertainty in the estimated pooled mean, not the range of predicted values for a new individual. For example, as the number of studies included in the meta-analysis increases, we would expect the confidence interval for the pooled mean to narrow, reflecting increased precision in the estimate. However, we would not expect the width of the estimated reference range to approach zero as the total sample size increases. Similarly, some have recently advocated for the reporting of a prediction interval for the mean or effect size of a new study when conducting meta-analyses in order to better describe between-study heterogeneity [65–68]. Riley et al. [67] describe a random effects meta-analysis example where there is a statistically significant pooled treatment effect, but the prediction interval for the treatment effect in a new study is [-0.79, 0.09]. They explain that the majority of the interval being below zero suggests that the treatment in question works in most settings, but that the small amount of the interval falling above zero indicates that the treatment may not be effective in some situations [67]. This example clearly illustrates how the confidence interval for the pooled mean does not necessarily represent the variation across study populations. However, the prediction interval for the mean of a new study still does not reflect the full variation on the individual participant level and would therefore not be suitable as a reference range either.

Figure 4.4: **Comparison of Intervals Estimated in Meta-Analysis** 95% Confidence interval for the pooled mean, 95% prediction interval for the mean of a new study, and estimated 95% reference range for $\hat{\mu} = 4, \hat{\sigma} = 1$, and $\hat{\tau} = 0.5$ and different within study sample size (n) and number of studies (N).



The differences in these intervals are illustrated in Figure 4.4, which shows the 95% confidence interval for the pooled mean, 95% prediction interval for a new study, and the estimated 95% reference range based on the same estimates of the pooled mean and within and between-study variances, but varying the numbers of studies included in the meta-analysis (N) and the number of individuals within each study (n). We can see that as the number of studies or number of participants within each study increases, the confidence interval for the pooled mean narrows. The prediction interval for a new study mean also narrows slightly, but this is due to greater perceived precision in the estimated parameters. Figure 4.4 also shows the estimated 95% reference ranges for each of these meta-analyses when using the frequentist method proposed by Siegel et al. [88]. This method does not incorporate uncertainty in the estimated parameters, so the width of the reference range does not change for different sample sizes. However, the Bayesian posterior predictive reference range interval, also proposed by Siegel et al. [88], can naturally incorporate the

uncertainty in the estimated parameters. Despite this difference, the estimated reference ranges in Figure 4.4 are still wider than other intervals. This is because they reflect both the estimated within-study and between-study variances, rather than only the between-study variance as does the prediction interval for the mean of a new study, or neither as does the confidence interval for the pooled mean.

4.10 Certainty about the estimated reference range

To be able to apply research evidence to patient care properly, evidence users (clinicians, patients, and guideline developers) need to know how certain or trustworthy is the evidence. Therefore, when a reference range is estimated, we need to consider applicability, risk of bias, heterogeneity and precision [92]. If possible, studies at high risk of bias (e.g., due to poor ascertainment of the measured laboratory test or because of a large proportion of patients lost to follow up) [93] could be excluded from the reference range estimation. If biased studies are included, the estimated between-study heterogeneity will reflect both true clinical differences in the study populations and heterogeneity caused by this bias [67]. If excluding these studies is not feasible, and we are left with a reference range estimated from studies at high risk of bias, certainty in this range will be low. If heterogeneity between the studies used to estimate the range was high and not explained by subgroup analyses, certainty will also be low. If the total sample size of included studies was small, the estimation of this range will also be imprecise and warrants lower certainty.

Furthermore, the Bayesian posterior predictive interval method for estimating the reference range incorporates the estimation uncertainty of parameters into the width of the interval, while the other methods do not. However, as previously mentioned, investigators primarily concerned about failing to diagnose a non-healthy patient may prefer not to use this method as the estimated reference range can contain greater than 95% of individual measurements. In this case, one could estimate the reference range using the frequentist or empirical method and then investigate uncertainty by comparing the results to the

Bayesian posterior predictive interval. If the Bayesian posterior predictive interval is considerably wider than the interval estimated using one of the other two methods, this would suggest a high degree of estimation uncertainty. However, there is still a need for further work to develop methods for quantifying the degree of uncertainty in the reference range limits estimated from a meta-analysis.

4.11 Discussion

This empirical application introduces the aggregate data approaches to estimating reference ranges proposed by Siegel et al. [88] and two one-step approaches using individual participant data. Overall, the results across all methods are similar in the clinical scenario, demonstrating that the aggregate data approaches with the corresponding log-transformation provide an adequate approximation to the results using the individual participant data. Each of the proposed methods is relatively easy to use. The Bayesian methods (both one and two-step) differ from the other methods in that the width of the estimated ranges increases with greater uncertainty. The frequentist and empirical approaches also do not require setting prior distributions for the model parameters and may be easier to implement in practice than the Bayesian methods. The frequentist methods can be implemented using existing software packages, while the empirical approach only uses simple formulas based on the aggregate data.

The assumptions used by each of the proposed methods should be considered when estimating the reference range, preferably by investigating distributional assumptions using IPD from at least 1-2 data sets, and further work is still needed to address situations where these assumptions are not met. However, each of the methods provides information about the variability of a measurement across healthy individuals beyond that provided by the pooled mean. The applied example using liver stiffness measurements also illustrates how these methods more accurately describe variation across healthy individuals than the confidence interval for the pooled mean or the prediction interval for the mean of a new

study. These methods are recommended to be used when estimating a reference range from a meta-analysis.

Chapter 5

Conclusion

5.1 Summary of major findings

This thesis proposed several methods for the meta-analysis of prevalence and normative data, two areas that have previously gained less attention in meta-analysis than the meta-analysis of randomized controlled trials or diagnostic test studies. Chapter 2 proposed a novel parameterization of a multivariate meta-analysis model for the joint meta-analysis of the prevalence of an overall outcome as well as several subtype outcomes. This parameterization accounted for the natural constraint that neither the underlying prevalence nor the observed counts of the subtypes can exceed those of the overall outcome. Simulation studies demonstrated that accounting for these natural constraints as well as the correlations between outcomes can reduce bias and increase precision, compared to both analyzing the outcomes univariately. This was also true when comparing this new parameterization to a multivariate model that does not account for these natural constraints. The simulations demonstrated these gains were largest in the presence of missing data, particularly when these data were missing at random. We hypothesize the reduced bias and increased precision are likely driven by the truncation of density in regions where the subtype outcomes would have higher prevalences than the overall outcome. These methods were demonstrated

using data from Markland et al. [19] on the prevalence of stress, urgency, or any type of urinary incontinence.

Chapter 3 proposed methods for estimating a reference range from a meta-analysis. No methodological guidance previously existed in the literature for estimation reference ranges using the results from multiple studies, particularly when only aggregate data are available. Siegel et al. [88] proposed three main approaches for estimating the meta-analysis that only require information on the sample size, observed mean, and standard deviation from each study included in the meta-analysis: a frequentist, a Bayesian, and an empirical approach. Simulation studies demonstrated that all three methods perform well in capturing the middle 95% of values when the true overall distribution was normal, the number of studies was relatively large (e.g. at least 20), and the between-study variance was relatively low compared to the overall variance (less than 30-50%). This was the case for both equal and unequal within-study variances. These methods were illustrated using two applied examples: pediatric waking time after sleep onset (WASO) and frontal subjective postural vertical measurements.

Finally, because Chapter 3 is written primarily for a statistical audience, Chapter 4 provides a guide aimed at a clinical and epidemiological audience describing how the three methods proposed in Chapter 3 can be used. Chapter 4 also extends these methods to the case where individual participant data are available. These methods are presented in the context of a clinical scenario about a patient at risk for liver fibrosis and who may undergo a non-invasive measure of liver stiffness for which there has previously been no established reference range. In this chapter, the concepts of heterogeneity, applicability, the target population, and the reference range's interpretation are explored more deeply. Finally, in the results for the clinical scenario, the estimated reference ranges using aggregate data are very similar to those using individual participant data, suggesting that the aggregate data approaches provide a valid alternative when individual participant data are not available.

5.2 Future research

The findings described in the previous section lead to many opportunities for future work. The first area stems from the work presented in Chapter 2 on reparameterizing multivariate meta-analysis models for binary outcomes to account for natural constraints in the data:

5.2.1 Bayesian Multivariate Meta-analysis of Serologic Test Accuracy: With Application to COVID-19

Serologic tests may measure the presence of multiple types of antibodies to determine whether a patient has a disease. For example, a recent systematic review and meta-analysis of COVID-19 antibody tests [94] assessed the sensitivity and specificity of different COVID-19 serologic tests when detecting IgG, IgM, or either type of antibody. In this case, the sensitivity when detecting either type of antibody cannot exceed the sensitivity when detecting either antibody individually, with the converse true for the specificities. Methods have previously been developed for the multivariate meta-analysis of diagnostic tests that jointly model sensitivity and specificity in order to account for the correlation between these two measures [4,5]. Hong et al. [42] also developed methods for meta-analysis in the case of both multiple treatments and multiple outcomes that could be easily applied to the diagnostic test setting. However, none of these methods account for the natural constraints described previously. We will extend the model proposed in Chapter 2 for multivariate meta-analysis of prevalences to the diagnostic test setting by also allowing for sensitivity and specificity to be jointly modeled in addition to the multiple antibody types.

5.2.2 Estimating the Reference Range from a Fixed Effects Meta-Analysis

The frequentist and Bayesian methods proposed in Chapter 3 for estimating the reference range from a meta-analysis are based on a random effects model. However, if only a small number of studies are included in the meta-analysis, it may be impossible to reliably estimate the between-study variance. Alternatively, the random effects normality assumption for the

study means may not appear reasonable in some cases. In these settings, one may prefer to estimate the reference range using a fixed effects model [64], which does not assume any particular relationship between study means. Cao et al. [95] recently proposed a novel method in addition to highlighting the empirical method proposed by Siegel et al. [88] for estimating the reference range from a meta-analysis using a fixed effects model. Further work is needed to establish when avoiding the assumptions made by the random effects model may be beneficial, as well as in which cases one of the two fixed effects methods proposed may be preferred over the other.

5.2.3 Incorporating Covariates when Estimating the Reference Range from a Meta-Analysis

When estimating the reference range, it is important to assess heterogeneity to establish whether it may be preferable to estimate separate reference ranges for subgroups, such as males and females. However, how these separate reference ranges are estimated requires further exploration. If only aggregate data are available and covariates are on the study level (e.g. study country, age groups), reference ranges can be estimated individually using separate models. However, if either the within or between-study variances can be assumed to be the same across subgroups, it may be more efficient to borrow information across groups by fitting a single meta-regression model. We will explore the relative benefits of different meta-regression models under a variety of settings as well as whether using model selection criteria (e.g. DIC in Bayesian settings) results in an appropriate choice of reference range model.

5.2.4 Estimating the Reference Range when Both Aggregate Data and IPD are Available

Chapter 3 proposes methods for estimating the reference range from a meta-analysis when only aggregate data are available and Chapter 4 extends these methods to cases where

IPD are available. However, if a meta-analysis contained some studies with IPD and some with only aggregate data, using the currently proposed methods, the studies with IPD would first need to be aggregated, then the three methods proposed in Chapter 3 could be applied. Future work is needed to develop methods for estimating the reference range that could be fit directly on both the aggregate data and IPD. This could potentially lead to greater efficiency and may allow for the use of individual patient level covariates while still incorporating information from the aggregated studies.

5.2.5 Nonparametric Estimation of the Reference Range from a Meta-Analysis

All of the proposed methods for estimating reference ranges [88, 95] make some parametric assumptions, either regarding the distributions of individuals' measurements within studies, the overall distribution across studies, or the distribution of the means of measurements across studies. The frequentist and Bayesian methods described in Chapters 3 and 4 make assumptions about all three. In some cases, these parametric assumptions may be deemed inappropriate, leading to the need to develop non-parametric methods for estimating the reference range.

5.2.6 Software

Finally, while the Web Supplement for the paper by Siegel et al. [88] presented in Chapter 3 provides code for implementing each of the proposed methods, developing software such as an R package or SAS macro will be important for increasing their use. The intended audience of Chapter 4 is comprised of clinicians and epidemiologists who may feel more comfortable implementing the proposed methods if they have access to user-friendly software. Eventually, this software will provide options for estimating the reference range from a meta-analysis using a variety of different methods depending on the modeling assumptions desired and availability of IPD.

References

[1] AB Haidich. Meta-analysis in medical research. *Hippokratia*, page 9, 2010.

[2] Irbaz Bin Riaz, Muhammad Shahzeb Khan, Haris Riaz, and Robert J. Goldberg. Disorganized Systematic Reviews and Meta-analyses: Time to Systematize the Conduct and Publication of These Study Overviews? *The American Journal of Medicine*, 129(3):339.e11–339.e18, March 2016.

[3] D Jackson, R Riley, and I R White. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, 30(20):2481–2498, 2011.

[4] Haitao Chu, Lei Nie, Stephen R. Cole, and Charles Poole. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: Alternative parameterizations and model selection. *Statistics in Medicine*, 28(18):2384–2399, 2009.

[5] H Chu, H Guo, and Y Zhou. Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. *Med Decis Making*, 30(4):499–508, 2010.

[6] Q Lian, J S Hodges, and H Chu. A bayesian hierarchical summary receiver operating characteristic model for network meta-analysis of diagnostic tests. *Journal of the American Statistical Association*, page in press, 2019.

[7] Fateh Bazerbachi, Samir Haffar, Zhen Wang, Joaquín Cabezas, Maria Teresa Arias-Loste, Javier Crespo, Sarwa Darwish-Murad, M. Arfan Ikram, John K. Olynyk, Eng

Gan, Salvatore Petta, Alessandra Berzuini, Daniele Prati, Victor de Lédinghen, Vincent W. Wong, Paolo Del Poggio, Norberto C. Chávez-Tapia, Yong-Peng Chen, Pin-Nan Cheng, Man-Fung Yuen, and Kausik Das. Range of Normal Liver Stiffness and Factors Associated With Increased Stiffness Measurements in Apparently Healthy Individuals. *Clinical Gastroenterology and Hepatology*, 17(1):54–64, January 2019.

[8] Poliana do Amaral Benfica, Larissa Tavares Aguiar, Sherindan Ayessa Ferreira de Brito, Luane Helena Nunes Bernardino, Luci Fuscaldi Teixeira-Salmela, and Christina Danielli Coelho de Morais Faria. Reference values for muscle strength: a systematic review with a descriptive meta-analysis. *Brazilian Journal of Physical Therapy*, 22(5):355–369, October 2018.

[9] Laila B. Conceição, Jussara A. O. Baggio, Suleimy C. Mazin, Dylan J. Edwards, and Taiza E. G. Santos. Normative data for human postural vertical: A systematic review and meta-analysis. *PLoS One; San Francisco*, 13(9):e0204122, September 2018.

[10] Jonathan M. Fawcett, Nichole Fairbrother, Emily J. Fawcett, and Ian R. White. A Bayesian multivariate approach to estimating the prevalence of a superordinate category of disorders. *International Journal of Methods in Psychiatric Research*, 27(4):e1742, 2018.

[11] Fibrinogen Studies Collaboration. Systematically missing confounders in individual participant data meta-analysis of observational cohort studies. *Statistics in Medicine*, 28(8):1218–1237, 2009.

[12] Barbara C. Galland, Michelle A. Short, Philip Terrill, Gabrielle Rigney, Jillian J. Haszard, Scott Coussens, Mistral Foster-Owens, and Sarah N. Biggs. Establishing normal values for pediatric nighttime sleep measured by actigraphy: a systematic review and meta-analysis. *Sleep*, 41(4), April 2018.

[13] RJ Rona, Keil T, C Summers, D Gislason, L Zuidmeer, E Sodergren, ST Sigurdardottir, T Lindner, K Goldhahn, J Dahlstrom, D McBride, and C Madsen. The prevalence of food allergy: A meta-analysis. *Journal of Allergy and Clinical Immunology*, 120(3):638–646, 2007.

[14] J G Williams, J P Higgins, and C E Brayne. Systematic review of prevalence studies of autism spectrum disorders. *Archives of Disease in Childhood*, 91(1):8–15, 2006.

[15] Daniel Campbell. Normative Data. In Fred R. Volkmar, editor, *Encyclopedia of Autism Spectrum Disorders*, pages 2062–2063. Springer New York, New York, NY, 2013.

[16] R D Riley. Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4):789–811, 2009.

[17] L Lin and H Chu. Bayesian multivariate meta-analysis of multiple factors. *Research Synthesis Methods*, 9(2):261–272, 2018.

[18] T A Trikalinos, D C Hoaglin, and C H Schmid. An empirical comparison of univariate and multivariate meta-analyses for categorical outcomes. *Statistics in Medicine*, 33(9):1441–1459, 2013.

[19] Alayne Markland, Haitao Chu, C. Neill Epperson, Jesse Nodora, David Shoham, Ariana Smith, Siobhan Sutcliffe, Mary Townsend, Jincheng Zhou, and Tamara Bavendam. Occupation and lower urinary tract symptoms in women: A rapid review and meta-analysis from the PLUS research consortium. *Neurourology and Urodynamics*, 37(8):2881–2892, 2018.

[20] Paul S. Horn and Amadeo J. Pesce. Reference intervals: an update. *Clinica Chimica Acta*, 334(1-2):5–23, August 2003.

[21] Paul S Horn, Amadeo J Pesce, and Bradley E Copeland. A robust approach to reference interval estimation and evaluation. *Clinical Chemistry*, 44(3):10, 1998.

[22] Lidia R. Arends, Zoltán Vokó, and Theo Stijnen. Combining multiple outcome measures in a meta-analysis: an application. *Statistics in Medicine*, 22(8):1335–1353, 2003.

[23] Sylwia Bujkiewicz, John R. Thompson, Alex J. Sutton, Nicola J. Cooper, Mark J. Harrison, Deborah P.M. Symmons, and Keith R. Abrams. Multivariate meta-analysis of mixed outcomes: a Bayesian approach. *Statistics in Medicine*, 32(22):3926–3943, September 2013.

[24] O Efthimiou, D Mavridis, R D Riley, A Cipriani, and G Salanti. Joint synthesis of multiple correlated outcomes in networks of interventions. *Biostatistics*, 16(1):84–97, 2014.

[25] O Efthimious, T P A Debray, G van Valkenhoef, S Trelle, K Panayidou, K G M Moons, J B Reitsma, A Shang, and G Salanti. GetReal in network meta-analysis: a review of the methodology. *Res. Syn. Meth.*, 7(3):236–263, 2016.

[26] X Ma, Q Lian, H Chu, J G Ibrahim, and Y Chen. A bayesian hierarchical model for network meta-analysis of multiple diagnostic tests. *Biostatistics*, 19(1):87–102, 2018.

[27] R D Riley, D Jackson, G Salanti, D L Burke, M Price, J Kirkham, and I R White. Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *BMJ*, 358(j3932), 2017.

[28] Hans C. Van Houwelingen, Koos H. Zwinderman, and Theo Stijnen. A bivariate approach to meta-analysis. *Statistics in Medicine*, 12(24):2273–2284, December 1993.

[29] Y Wei and J P T Higgins. Bayesian multivariate meta-analysis with multiple outcomes. *Statistics in Medicine*, 32(17):2911–2934, 2013.

[30] J Zhang, B P Carlin, J D Neaton, G G Soon, L Nie, R Kane, B A Virnig, and H Chu. Network meta-analysis of randomized clinical trials: Reporting the proper summaries. *Clinical Trials*, 11(2):246–262, 2014.

[31] J J Kirkham, R D Riley, and P R Williamson. A multivariate meta-analysis approach for reducing bias in systematic reviews. *Statistics in Medicine*, 31(20):2179–2195, 2012.

[32] Donald R. Williams and Paul-Christian Bürkner. Effects of intranasal oxytocin on symptoms of schizophrenia: A multivariate bayesian meta-analysis. *Psychoneuroendocrinology*, 75:141–151, 2018/10/25 2017.

[33] Yong Chen, Yulun Liu, Haitao Chu, Mei-Ling Ting Lee, and Christopher H. Schmid. A simple and robust method for multivariate meta-analysis of diagnostic test accuracy. *Statistics in medicine*, 36(1):105–121, January 2017.

[34] Oliver Kuss, Annika Hoyer, and Alexander Solms. Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas. *Statistics in Medicine*, 33(1):17–30, January 2014.

[35] J B Reitsma, A S Glas, A W S Rutjes, R J P M Scholten, P M Bossuyt, and A H Zwinderman. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol.*, 58(10):982–990, 2005.

[36] F S Tonin, L M Steimbach, A M Mendes, HH Borba, R Pontarolo, and F Fernandez-Llimos. Mapping the characteristics of network meta- analyses on drug therapy: A systematic review. *PLoS ONE*, 13(4):e0196644, 2018.

[37] Joshua D. Niforatos, Matt Weaver, and Michael E. Johansen. Assessment of Publication Trends of Systematic Reviews and Randomized Clinical Trials, 1995 to 2017. *JAMA Internal Medicine*, July 2019.

[38] Jan J. Barendregt, Suhail A. Doi, Yong Yi Lee, Rosana E. Norman, and Theo Vos. Meta-analysis of prevalence. *J Epidemiol Community Health*, 67(11):974–978, November 2013.

[39] RD Riley, KR Abrams, P C Lambert, A J Sutton, and J R Thompson. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine*, 26(1):78–97, 2007.

[40] Bernard L. Harlow, Tamara G. Bavendam, Mary H. Palmer, Linda Brubaker, Kathryn L. Burgio, Emily S. Lukacz, Janis M. Miller, Elizabeth R. Mueller, Diane K. Newman, Leslie M. Rickey, Siobhan Sutcliffe, Denise Simons-Morton, and On behalf of The PLUS Research Consortium. The Prevention of Lower Urinary Tract Symptoms (PLUS) Research Consortium: A Transdisciplinary Approach Toward Promoting Bladder Health and Preventing Lower Urinary Tract Symptoms in Women Across the Life Course. *Journal of Women's Health*, 27(3):283–289, March 2018.

[41] Teresa C. Smith, David J. Spiegelhalter, and Andrew Thomas. Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine*, 14(24):2685–2699, 1995.

[42] Hwanhee Hong, Haitao Chu, Jing Zhang, and Bradley P. Carlin. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research Synthesis Methods*, 7(1):6–22, 2016.

[43] Georgia Salanti, Julian PT Higgins, Ae Ades, and John PA Ioannidis. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research*, 17(3):279–301, June 2008.

[44] A Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

[45] A Alvarez, J Niemi, and M Simpson. Bayesian inference for a covariance matrix. 26th Annual Conference Proceedings, Annual Conference on Applied Statistics in Agriculture, pages 71–82, 2014.

[46] J Barnard, R McCulloch, and X Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1311, 2000.

[47] G Lu and A E Ades. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics*, 10(4):792–805, 2009.

[48] LF Lin, J Zhang, JS Hodges, and HT Chu. Performing arm-based network meta-analysis in R with the pcnetmeta package. *Journal Of Statistical Software*, 80(5):1–25, 2017.

[49] M Plummer. Jags: A program for analysis of bayesian graphical models using Gibbs sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing, 2003.

[50] R Core Team. *R: A Language and Environment fxor Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[51] Martyn Plummer. *rjags: Bayesian Graphical Models using MCMC*, 2016. R package version 4-6.

[52] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006.

[53] Richard W Bohannon. Reference Values for the Timed Up and Go Test: A Descriptive Meta-Analysis. *Journal of Geriatric Physical Therapy*, 29:64–68, 2006.

[54] Richard W. Bohannon and A. Williams Andrews. Normal walking speed: a descriptive meta-analysis. *Physiotherapy*, 97(3):182–189, September 2011.

[55] Richard M. Dodds, Holly E. Syddall, Rachel Cooper, Diana Kuh, Cyrus Cooper, and Avan Aihie Sayer. Global variation in grip strength: a systematic review and meta-analysis of normative data. *Age and Ageing*, 45(2):209–216, March 2016.

[56] Balázs Németh, Zénó Ajtay, László Hejjel, Tamás Ferenci, Zoltán Ábrám, Edit Murányi, and István Kiss. The issue of plasma asymmetric dimethylarginine reference range - A systematic review and meta-analysis. *PLOS ONE*, 12(5):e0177493, May 2017.

[57] Juxian Tang, Yihui Lin, Huachao Mai, Yiping Luo, Renwei Huang, Qi Chen, and Duan Xiao. Meta-analysis of reference values of haemostatic markers during pregnancy and childbirth. *Taiwanese Journal of Obstetrics and Gynecology*, 58(1):29–35, January 2019.

[58] Jean F. Wyman, Jincheng Zhou, D. Y. LaCoursiere, Alayne D. Markland, Elizabeth R. Mueller, Laura Simon, Ann Stapleton, Carolyn R. T. Stoll, Haitao Chu, and Siobhan Sutcliffe. Normative noninvasive bladder function measurements in healthy women: A systematic review and meta-analysis. *Neurourology and Urodynamics*, 39(2):507–522, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nau.24265.

[59] Haoming Xu, Maira Fonseca, Zachary Wolner, Esther Chung, Xinyuan Wu, Shamir Geller, Stephen W. Dusza, Antonio P. DeRosa, Ashfaq A. Marghoob, Klaus J. Busam, Allan C. Halpern, and Michael A. Marchetti. Reference values for skin microanatomy: A systematic review and meta-analysis of exÂ vivo studies. *Journal of the American Academy of Dermatology*, 77(6):1133–1144.e4, December 2017.

[60] Ali Reza Khoshdel, Ammarin Thakkinstian, Shane L Carney, and John Attia. Estimation of an age-specific reference interval for pulse wave velocity: a meta-analysis:. *Journal of Hypertension*, 24(7):1231–1237, July 2006.

[61] Philip T. Levy, Aliza Machefsky, Aura A. Sanchez, Meghna D. Patel, Sarah Rogal, Susan Fowler, Lauren Yaeger, Angela Hardi, Mark R. Holland, Aaron Hamvas, and Gautam K. Singh. Reference Ranges of Left Ventricular Strain Measures by Two-Dimensional Speckle-Tracking Echocardiography in Children: A Systematic Review and Meta-Analysis. *Journal of the American Society of Echocardiography*, 29(3):209–225.e6, March 2016.

[62] Jan A. Staessen, Robert H. Fagard, Paul J. Lijnen, Lutgarde Thijs, Roger Van Hoof, and Antoon K. Amery. Mean and range of the ambulatory pressure in normotensive subjects from a meta-analysis of 23 studies. *The American Journal of Cardiology*, 67(8):723–727, April 1991.

[63] Allison A Venner, Patricia K Doyle-Baker, Martha E Lyon, and Tak S Fung. A meta-analysis of leptin reference ranges in the healthy paediatric prepubertal population. *Annals of Clinical Biochemistry*, 46(1):65–72, January 2009. Publisher: SAGE Publications.

[64] Nan M. Laird and Frederick Mosteller. Some Statistical Methods for Combining Experimental Results. *International Journal of Technology Assessment in Health Care*, 6(01):5–30, January 1990.

[65] Joanna IntHout, John P. A. Ioannidis, Maroeska M. Rovers, and Jelle J. Goeman. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*, 6(7):e010247, July 2016.

[66] Lifeng Lin. Use of Prediction Intervals in Network Meta-analysis. *JAMA Network Open*, 2(8):e199735, August 2019.

[67] Richard D. Riley, Julian P. T. Higgins, and Jonathan J. Deeks. Interpretation of random effects meta-analyses. *BMJ*, 342:d549, February 2011.

[68] Chia-Chun Wang and Wen-Chung Lee. A simple method to estimate prediction intervals and predictive distributions: Summarizing meta-analyses beyond means and confidence intervals. *Research Synthesis Methods*, 10(2):255–266, 2019.

[69] Ralf Bender, Tim Friede, Armin Koch, Oliver Kuss, Peter Schlattmann, Guido Schwarzer, and Guido Skipka. Methods for evidence synthesis in the case of very few studies. *Research Synthesis Methods*, 9(3):382–392, 2018.

[70] Michael Borenstein, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2):97–111, 2010.

[71] Kenneth Rice, Julian P. T. Higgins, and Thomas Lumley. A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(1):205–227, January 2018.

[72] Ana C. Alba, Paul E. Alexander, Joanne Chang, John MacIsaac, Samantha DeFry, and Gordon H. Guyatt. High statistical heterogeneity is more frequent in meta-analysis of continuous than binary outcomes. *Journal of Clinical Epidemiology*, 70:129–135, February 2016.

[73] Julian P T Higgins, Simon G Thompson, and David J Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, 172(1):137–159, January 2009.

[74] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, September 1986.

[75] Guido Schwarzer. meta: An R package for meta-analysis. *R News*, 7(3):40–45, 2007.

[76] John E. Cornell, Cynthia D. Mulrow, Russell Localio, Catharine B. Stack, Anne R. Meibohm, Eliseo Guallar, and Steven N. Goodman. Random-Effects Meta-analysis of

Inconsistent Effects: A Time for Change. *Annals of Internal Medicine*, 160(4):267–270, February 2014.

[77] Areti Angeliki Veroniki, Dan Jackson, Wolfgang Viechtbauer, Ralf Bender, Jack Bowden, Guido Knapp, Oliver Kuss, Julian PT Higgins, Dean Langan, and Georgia Salanti. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7(1):55–79, March 2016.

[78] Wolfgang Viechtbauer. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics*, 30(3):261–293, September 2005.

[79] Hussam Alshraideh, Hazem Smadi, Jalal Abo-Taha, and Obaidah Alomari. Reference Range Estimation: Accounting for Measurement System Errors. *Quality and Reliability Engineering International*, 32(3):901–908, 2016.

[80] Robert G. Hoffmann. Statistics in the Practice of Medicine. *JAMA: The Journal of the American Medical Association*, 185(11):864, September 1963.

[81] Thomas W. D. Möbius. *metagen: Inference in Meta Analysis and Meta Regression*, 2014. R package version 1.0.

[82] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[83] Royette Tavernier, Sungsub B Choo, Kathryn Grant, and Emma K Adam. Daily Affective Experiences Predict Objective Sleep Outcomes among Adolescents. *Journal of sleep research*, 25(1):62–69, February 2016.

[84] Jussara A. O. Baggio, Suleimy S. C. Mazin, Frederico F. Alessio-Alves, Camila G. C. Barros, Antonio A. O. Carneiro, JoÃ£o P. Leite, Octavio M. Pontes-Neto, and Taiza

E. G. Santos-Pontelli. Verticality Perceptions Associate with Postural Control and Functionality in Stroke Patients. *PLoS ONE*, 11(3):e0150754, March 2016.

[85] Seymour Geisser. *Predictive inference : an introduction*. Monographs on statistics and applied probability (Series) ; 55. Chapman & Hall, New York, 1993.

[86] Faraz Pathan, Nicholas D'Elia, Mark Nolan, Thomas Marwick, and Kazuaki Negishi. NORMAL RANGES OF LEFT ATRIAL STRAIN BY SPECKLE TRACKING ECHOCARDIOGRAPHY: A SYSTEMATIC REVIEW AND META-ANALYSIS OF 1,789 HEALTHY SUBJECTS. *Journal of the American College of Cardiology*, 67(13):1582, April 2016.

[87] Darrick K. Li, Muhammad Rehan Khan, Zhen Wang, Voranush Chongsrisawat, Panida Swangsak, Ulrike Teufel-Schäfer, Guido Engelmann, Imeke Goldschmidt, Ulrich Baumann, Daisuke Tokuhara, Yuki Cho, Marion Rowland, Anders B. Mjelle, Grant A. Ramm, Peter J. Lewindon, Peter Witters, David Cassiman, Ioana M. Ciuca, Larry D. Prokop, Samir Haffar, Kathleen E. Corey, M. H. Murad, Katryn N. Furuya, and Fateh Bazerbachi. Normal liver stiffness and influencing factors in healthy children: An individual participant data meta-analysis. *Liver International*, 40(11):2602–2611, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/liv.14658.

[88] Lianne Siegel, M. Hassan Murad, and Haitao Chu. Estimating the reference range from a meta-analysis. *Research Synthesis Methods*, 12(2):148–160, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.1442.

[89] Varun Takyar, Anand Nath, Andrea Beri, Ahmed M. Gharib, and Yaron Rotman. How healthy are the "Healthy volunteers"? Penetrance of NAFLD in the biomedical research volunteer pool. *Hepatology*, 66(3):825–833, 2017. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hep.29247.

[90] W. L. Baker, C. Michael White, J. C. Cappelleri, J. Kluger, and C. I. Coleman. Understanding heterogeneity in meta-analysis: the role of meta-regression. *International Journal of Clinical Practice*, 63(10):1426–1434, 2009. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1742-1241.2009.02168.x.

[91] Kengo Nagashima, Hisashi Noma, and Toshi A Furukawa. Prediction intervals for random-effects meta-analysis: A confidence distribution approach. *Statistical Methods in Medical Research*, 28(6):1689–1702, June 2019. Publisher: SAGE Publications Ltd STM.

[92] M. Hassan Murad. Clinical Practice Guidelines: A Primer on Development and Dissemination. *Mayo Clinic Proceedings*, 92(3):423–433, March 2017.

[93] Mohammad Hassan Murad, Shahnaz Sultan, Samir Haffar, and Fateh Bazerbachi. Methodological quality and synthesis of case series and case reports. *BMJ Evidence-Based Medicine*, 23(2):60–63, April 2018.

[94] Mayara Lisboa Bastos, Gamuchirai Tavaziva, Syed Kunal Abidi, Jonathon R Campbell, Louis-Patrick Haraoui, James C Johnston, Zhiyi Lan, Stephanie Law, Emily MacLean, Anete Trajman, Dick Menzies, Andrea Benedetti, and Faiz Ahmad Khan. Diagnostic accuracy of serological tests for covid-19: systematic review and meta-analysis. *BMJ*, page m2516, July 2020.

[95] Wenhao Cao, Lianne Siegel, Jincheng Zhou, Motao Zhu, Tiejun Tong, Yong Chen, and Haitao Chu. Estimating the Reference Interval from a Fixed Effects Meta-Analysis. *Research Synthesis Methods*, in press.

[96] George Casella. *Statistical inference*. Thomson Learning, Australia ; Pacific Grove, Calif., 2nd ed.. edition, 2002.

Appendix A

Supplementary Materials for "A Bayesian Multivariate Meta-Analysis of Prevalence Data"

A.1 Marginal Event Rate for Subtypes

In this section we derive Equation (2.11), which describes how to find the population averaged prevalence (π_j) for each subtype j.

This is given by:

$$\pi_j = E[\pi_{ij}] = E[\pi_{i0} p_{ij}] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi(\mu_0 + \sigma_0 z_0) \Phi(\mu_j + \sigma_j^* z_j) \phi(z_0, z_j) dz_0 dz_j, \qquad \text{(A.1)}$$

where $\phi(z_0, z_j)$ denotes the probability density function of a bivariate standard normal distribution.

This is equivalent to:

$$\begin{aligned} &E[P(Y_1 \leq \mu_0 + \sigma_0 Z_0)P(Y_2 \leq \mu_j + \sigma_j^* Z_j)] \\ =&E[P(Y_1 - \sigma_0 Z_0 \leq \mu_0)P(Y_2 - \sigma_j^* Z_j) \leq \mu_j^*], \end{aligned} \tag{A.2}$$

where Y_1 and Y_2 are two independent standard normal random variables. Because $Y_1 - \sigma_0 z_0 \sim N(0, 1+\sigma_0^2)$ and $Y_2 - \sigma_j^* z_j \sim N(0, 1+\sigma_j^{*2})$, we can do the transformation: $T_1 = \frac{Y_1 - \sigma_0 Z_0}{\sqrt{1+\sigma_0^2}}$ and $T_2 = \frac{Y_2 - \sigma_j^* Z_j}{\sqrt{1+\sigma_j^{*2}}}$, where T_1 and T_2 are both distributed $N(0,1)$ with correlation ρ. Therefore, this is equal to:

$$E\left[P\left(T_1 \leq \frac{\mu_0}{\sqrt{1+\sigma_0^2}}\right) P\left(T_2 \leq \frac{\mu_j}{\sqrt{1+\sigma_j^{*2}}}\right)\right] \tag{A.3}$$

Since Z_0 and Z_j are distributed bivariate standard normal, we can see that (T_1, T_2) are distributed bivariate normal with means (0,0), variances (1,1) and some $Cov[T_1, T_2]$.

We can then solve for this covariance:

$$\begin{aligned} Cov[T_1, T_2] &= Cov\left[\frac{Y_1 - \sigma_0 Z_0}{\sqrt{1+\sigma_0^2}}, \frac{Y_2 - \sigma_j^* Z_j}{\sqrt{1+\sigma_j^{*2}}}\right] \\ &= \frac{1}{\sqrt{1+\sigma_0^2}\sqrt{1+\sigma_j^{*2}}} Cov\left[Y_1 - \sigma_0 Z_0, Y_2 - \sigma_j^* Z_j\right] \\ &= \frac{1}{\sqrt{1+\sigma_0^2}\sqrt{1+\sigma_j^2}} Cov\left[\sigma_0 Z_0, \sigma_j^* Z_j\right] \end{aligned} \tag{A.4}$$

$Cov\left[\sigma_0 Z_0, \sigma_j^* Z_j\right]$ is equal to $\mathbf{\Sigma^*}_{1,j+1}$, which we are already estimating. Therefore, we can estimate π_j using:

$$\pi_j = P\left(X < \frac{\mu_0}{\sqrt{1+\sigma_0^2}}, Y < \frac{\mu_j}{\sqrt{1+\sigma_j^{*2}}}\right) \tag{A.5}$$

where X and Y are distributed bivariate normal with means = $(0, 0)$, variances = (1, 1), and covariance: $Cov\left[X, Y\right] = \frac{1}{\sqrt{1+\sigma_0^2}\sqrt{1+\sigma_j^{*2}}}\mathbf{\Sigma^*}_{1,j+1}$.

A.2 Tables and Figures

Figure A.1: **Posterior Density Plot** Posterior prevalence density plots for overall UI and subtypes (SUI, UUI) for univariate model, multivariate model, and new parameterization

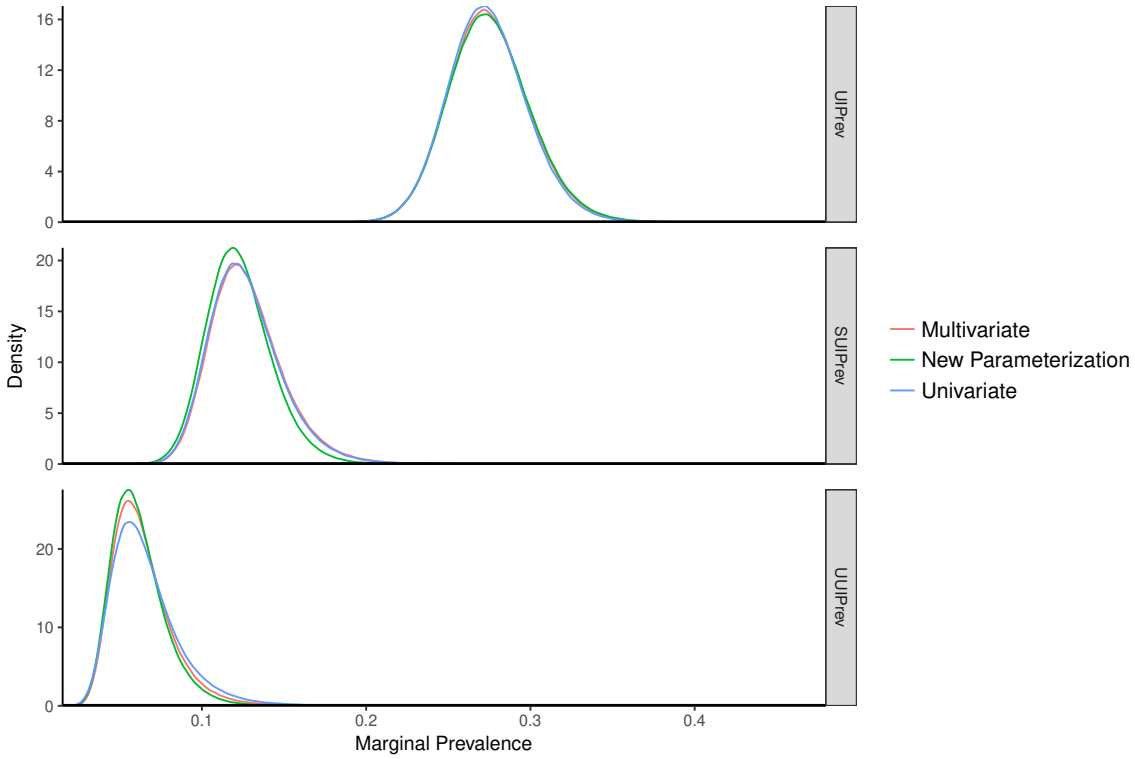


Table A.1: **Case Study Data** Author, publication year, sample size (N), and counts for any urinary incontinence (UIPrev), stress incontience (SUIPrev), or urgency incontinence (UUIPrev)

	Author	PubYear	N	UIPrev	SUIPrev	UUIPrev
1	Araki	2005	3734	624	456	76
2	Azuma	2008	975	220	42	9
3	Bailey	2010	200	55	28	28
4	Bo	2011	685	181	47	10
5	Buschbaum	2002	149	75	23	18
6	Davis	1999	563	175		
7	Fischer	1999	274	72	51	7
8	Fitzgerald	2000	1113	234	197	
9	Fitzgerald	2002	269	78	20	1
10	Fultz	2005	3364	1480	148	548
11	Liao	2007	445	120	62	8
12	Liao	2009	907	82	25	14
13	Nygaard	1997	791	127	23	17
14	Liu	2014	5433	1684		
15	Kaya	2016	281	51	25	14
16	Kim	2016	5928	445		
17	Hart	1999	1113	234		
18	Lam	1992	2631	510		
19	Palmer	2015	113	60		
20	Pierce	2017	2907	930		
21	Peyrat	2002	1700	357	161	21
22	Saadoun	2006	2640	554		
23	Sexton	2009	2820	722		
24	Singh	2013	3000	657	484	62
25	Wan	2016	636	297	122	73
26	Zhang	2013	1070	482	349	225

Table A.2: **(N = 30, n_i = 500, MCAR)** Bias, 95% credible interval width (CIW) and coverage probability (Cov.) for univariate model, original multivariate model, and new parameterization across 2000 simulations containing 30 studies, where $n_i = 500$, the true prevalences are (0.3, 0.15, 0.05), and data are missing completely at random (MCAR).

		Overall			**Subtype 1**			**Subtype 2**		
		Bias	CIW	Cov.	Bias	CIW	Cov.	Bias	CIW	Cov.
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0$	Univariate	0.003	0.118	0.95	0.009	0.121	0.96	0.007	0.070	0.95
	Original	0.004	0.120	0.96	0.007	0.098	0.96	0.008	0.064	0.95
	New Param.	0.004	0.121	0.96	0.002	0.093	0.96	0.005	0.057	0.96
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0.4$	Univariate	0.003	0.118	0.96	0.010	0.138	0.96	0.009	0.083	0.97
	Original	0.003	0.119	0.96	0.007	0.104	0.96	0.007	0.067	0.96
	New Param.	0.004	0.121	0.96	0.003	0.100	0.96	0.005	0.061	0.96
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0.8$	Univariate	0.003	0.118	0.96	0.011	0.153	0.96	0.011	0.094	0.96
	Original	0.003	0.118	0.96	0.006	0.106	0.95	0.006	0.063	0.95
	New Param.	0.003	0.119	0.96	0.004	0.103	0.95	0.005	0.059	0.95
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0$	Univariate	0.003	0.118	0.95	0.013	0.159	0.97	0.017	0.122	0.97
	Original	0.004	0.121	0.96	0.015	0.147	0.95	0.020	0.126	0.96
	New Param.	0.005	0.121	0.96	0.003	0.119	0.96	0.009	0.085	0.96
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0.4$	Univariate	0.003	0.118	0.96	0.016	0.181	0.97	0.020	0.139	0.96
	Original	0.004	0.120	0.96	0.014	0.153	0.97	0.020	0.127	0.95
	New Param.	0.004	0.121	0.96	0.004	0.126	0.96	0.011	0.093	0.95
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0.8$	Univariate	0.003	0.118	0.96	0.017	0.198	0.97	0.023	0.155	0.96
	Original	0.003	0.119	0.96	0.010	0.146	0.97	0.015	0.111	0.96
	New Param.	0.003	0.120	0.96	0.005	0.129	0.95	0.010	0.092	0.97

Table A.3: **(N = 30, n_i = 500, MAR)** Bias, 95% credible interval width (CIW) and coverage probability (Cov.) for univariate model, original multivariate model, and new parameterization, across 2000 simulations containing 30 studies, where $n_i = 500$, the true prevalences are (0.3, 0.15, 0.05), and data are missing at random (MAR).

		Overall			Subtype 1			Subtype 2		
		Bias	CIW	Cov.	Bias	CIW	Cov.	Bias	CIW	Cov.
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0$	Univariate	0.003	0.118	0.954	0.009	0.121	0.963	0.007	0.070	0.953
	Original	0.004	0.120	0.959	0.007	0.098	0.958	0.008	0.064	0.949
	New Param.	0.004	0.121	0.963	0.002	0.093	0.962	0.005	0.057	0.960
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0.4$	Univariate	0.004	0.118	0.960	0.078	0.143	0.309	0.036	0.092	0.503
	Original	0.004	0.119	0.964	0.014	0.114	0.946	0.012	0.074	0.942
	New Param.	0.005	0.121	0.967	0.005	0.096	0.956	0.006	0.056	0.959
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0.8$	Univariate	0.003	0.118	0.960	0.092	0.154	0.234	0.044	0.098	0.381
	Original	0.004	0.118	0.961	0.009	0.106	0.953	0.008	0.061	0.945
	New Param.	0.004	0.119	0.963	0.007	0.101	0.953	0.006	0.055	0.948
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0$	Univariate	0.004	0.118	0.954	0.069	0.187	0.630	0.039	0.150	0.804
	Original	0.005	0.121	0.962	0.036	0.201	0.951	0.038	0.180	0.923
	New Param.	0.005	0.121	0.961	0.003	0.111	0.965	0.011	0.081	0.954
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0.4$	Univariate	0.004	0.118	0.963	0.088	0.199	0.487	0.051	0.161	0.661
	Original	0.005	0.120	0.966	0.024	0.173	0.959	0.028	0.147	0.934
	New Param.	0.005	0.121	0.965	0.006	0.120	0.961	0.012	0.084	0.948
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0.8$	Univariate	0.003	0.118	0.957	0.111	0.205	0.322	0.062	0.164	0.510
	Original	0.004	0.119	0.959	0.014	0.139	0.945	0.014	0.098	0.947
	New Param.	0.004	0.120	0.962	0.009	0.126	0.947	0.010	0.081	0.945

Table A.4: **(N = 10, n_i = 100, MCAR)** Bias, 95% credible interval width (CIW) and coverage probability (Cov.) for univariate model, original multivariate model, and new parameterization across 2000 simulations containing 10 studies, where $n_i = 100$, the true prevalences are (0.3, 0.15, 0.05), and data are missing completely at random (MCAR).

		Overall			**Subtype 1**			**Subtype 2**		
		Bias	CIW	Cov.	Bias	CIW	Cov.	Bias	CIW	Cov.
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0$	Univariate	0.009	0.226	0.96	0.040	0.332	0.97	0.052	0.314	0.96
	Original	0.012	0.237	0.97	0.033	0.278	0.97	0.053	0.289	0.96
	New Param.	0.013	0.241	0.97	0.007	0.198	0.98	0.022	0.162	0.98
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0.4$	Univariate	0.008	0.225	0.96	0.046	0.363	0.98	0.059	0.340	0.97
	Original	0.011	0.235	0.97	0.036	0.299	0.97	0.054	0.297	0.96
	New Param.	0.012	0.240	0.97	0.008	0.210	0.97	0.023	0.170	0.97
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0.8$	Univariate	0.008	0.225	0.96	0.048	0.378	0.97	0.067	0.364	0.96
	Original	0.009	0.232	0.97	0.034	0.298	0.97	0.055	0.297	0.97
	New Param.	0.011	0.235	0.97	0.010	0.214	0.97	0.025	0.176	0.97
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0$	Univariate	0.007	0.225	0.95	0.051	0.389	0.97	0.077	0.397	0.96
	Original	0.011	0.238	0.97	0.052	0.358	0.96	0.086	0.384	0.94
	New Param.	0.011	0.240	0.97	0.004	0.226	0.97	0.028	0.192	0.96
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0.4$	Univariate	0.006	0.224	0.96	0.058	0.421	0.98	0.081	0.408	0.96
	Original	0.009	0.234	0.97	0.054	0.374	0.97	0.083	0.382	0.95
	New Param.	0.010	0.237	0.97	0.006	0.238	0.97	0.029	0.200	0.96
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0.8$	Univariate	0.006	0.223	0.96	0.060	0.436	0.97	0.082	0.413	0.97
	Original	0.008	0.229	0.96	0.047	0.364	0.97	0.073	0.360	0.96
	New Param.	0.008	0.231	0.96	0.007	0.244	0.96	0.030	0.207	0.96

Table A.5: **(N = 10, n_i = 100, MAR)** Bias, 95% credible interval width (CIW) and coverage probability (Cov.) for univariate model, original multivariate model, and new parameterization across 2000 simulations containing 10 studies, where $n_i = 100$, the true prevalences are (0.3, 0.15, 0.05), and data are missing at random (MAR).

		Overall			**Subtype 1**			**Subtype 2**		
		Bias	CIW	Cov.	Bias	CIW	Cov.	Bias	CIW	Cov.
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0$	Univariate	0.009	0.226	0.96	0.089	0.342	0.83	0.070	0.332	0.86
	Original	0.012	0.238	0.97	0.060	0.336	0.95	0.072	0.344	0.93
	New Param.	0.013	0.240	0.97	0.006	0.185	0.98	0.022	0.146	0.97
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0.4$	Univariate	0.009	0.225	0.96	0.101	0.356	0.77	0.079	0.341	0.80
	Original	0.012	0.236	0.97	0.058	0.332	0.95	0.066	0.326	0.93
	New Param.	0.013	0.239	0.97	0.012	0.196	0.98	0.024	0.150	0.96
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0.8$	Univariate	0.009	0.225	0.96	0.112	0.364	0.73	0.084	0.338	0.73
	Original	0.011	0.232	0.96	0.052	0.312	0.95	0.056	0.295	0.92
	New Param.	0.012	0.237	0.97	0.017	0.202	0.97	0.025	0.150	0.94
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0$	Univariate	0.009	0.226	0.96	0.103	0.417	0.86	0.100	0.432	0.90
	Original	0.013	0.238	0.97	0.090	0.438	0.94	0.117	0.465	0.92
	New Param.	0.013	0.238	0.97	0.006	0.210	0.98	0.029	0.175	0.96
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0.4$	Univariate	0.009	0.225	0.96	0.120	0.428	0.81	0.112	0.440	0.84
	Original	0.013	0.236	0.97	0.084	0.425	0.94	0.109	0.449	0.92
	New Param.	0.013	0.237	0.97	0.014	0.224	0.97	0.034	0.184	0.95
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0.8$	Univariate	0.009	0.224	0.95	0.137	0.430	0.74	0.117	0.432	0.77
	Original	0.011	0.232	0.96	0.069	0.385	0.94	0.082	0.386	0.92
	New Param.	0.012	0.234	0.96	0.021	0.234	0.95	0.034	0.189	0.93

Table A.6: **(N = 30, n_i = 100, No Missing Data)** Bias, 95% credible interval width (CIW) and coverage probability (Cov.) for univariate model, original multivariate model, and new parameterization across 2000 simulations containing 30 studies, where $n_i = 100$, the true prevalences are (0.3, 0.15, 0.05), and the data are fully observed.

		Overall			**Subtype 1**			**Subtype 2**		
		Bias	CIW	Cov.	Bias	CIW	Cov.	Bias	CIW	Cov.
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0$	Univariate	0.003	0.121	0.95	0.004	0.081	0.96	0.003	0.044	0.95
	Original	0.003	0.123	0.96	0.005	0.083	0.96	0.004	0.045	0.95
	New Param.	0.004	0.124	0.96	0.002	0.081	0.96	0.003	0.043	0.96
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0.4$	Univariate	0.003	0.121	0.95	0.005	0.092	0.96	0.004	0.051	0.96
	Original	0.003	0.123	0.96	0.006	0.094	0.96	0.005	0.052	0.96
	New Param.	0.004	0.124	0.96	0.004	0.092	0.96	0.004	0.050	0.96
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0.8$	Univariate	0.003	0.121	0.96	0.005	0.102	0.96	0.005	0.058	0.96
	Original	0.006	0.134	0.95	0.009	0.119	0.95	0.008	0.073	0.95
	New Param.	0.003	0.122	0.96	0.005	0.100	0.96	0.004	0.054	0.95
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0$	Univariate	0.003	0.121	0.96	0.006	0.104	0.96	0.007	0.071	0.96
	Original	0.004	0.124	0.96	0.008	0.106	0.96	0.009	0.075	0.96
	New Param.	0.005	0.125	0.96	0.003	0.099	0.96	0.005	0.063	0.96
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0.4$	Univariate	0.003	0.121	0.96	0.007	0.118	0.96	0.009	0.082	0.96
	Original	0.004	0.123	0.96	0.008	0.119	0.97	0.010	0.084	0.96
	New Param.	0.005	0.125	0.96	0.005	0.112	0.96	0.007	0.073	0.96
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0.8$	Univariate	0.003	0.121	0.96	0.008	0.131	0.96	0.010	0.090	0.96
	Original	0.003	0.122	0.96	0.008	0.130	0.96	0.010	0.088	0.96
	New Param.	0.003	0.122	0.96	0.006	0.122	0.95	0.008	0.079	0.96

Table A.7: **(N = 10, n_i = 100, No Missing Data)** Bias, 95% credible interval width (CIW) and coverage probability (Cov.) for univariate model, original multivariate model, and new parameterization across 2000 simulations containing 10 studies, where $n_i = 100$, the true prevalences are (0.3, 0.15, 0.05), and the data are fully observed.

		Overall			**Subtype 1**			**Subtype 2**		
		Bias	CIW	Cov.	Bias	CIW	Cov.	Bias	CIW	Cov.
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0$	Univariate	0.008	0.224	0.96	0.012	0.165	0.96	0.013	0.113	0.96
	Original	0.011	0.235	0.97	0.016	0.175	0.96	0.016	0.125	0.95
	New Param.	0.013	0.244	0.97	0.007	0.166	0.97	0.010	0.104	0.96
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0.4$	Univariate	0.007	0.223	0.95	0.015	0.185	0.95	0.016	0.133	0.96
	Original	0.009	0.231	0.96	0.017	0.193	0.96	0.019	0.141	0.96
	New Param.	0.012	0.240	0.97	0.011	0.183	0.96	0.013	0.118	0.96
$\boldsymbol{\sigma} = (0.5, 0.5, 0.5), \boldsymbol{\rho} = 0.8$	Univariate	0.007	0.222	0.96	0.016	0.203	0.96	0.020	0.153	0.96
	Original	0.009	0.231	0.96	0.019	0.212	0.97	0.021	0.155	0.96
	New Param.	0.010	0.233	0.97	0.014	0.196	0.96	0.016	0.131	0.96
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0$	Univariate	0.007	0.224	0.96	0.017	0.209	0.97	0.029	0.194	0.97
	Original	0.011	0.241	0.97	0.024	0.226	0.97	0.040	0.221	0.95
	New Param.	0.013	0.246	0.97	0.005	0.194	0.97	0.018	0.147	0.97
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0.4$	Univariate	0.006	0.224	0.96	0.020	0.236	0.97	0.034	0.216	0.97
	Original	0.009	0.236	0.97	0.026	0.247	0.97	0.042	0.233	0.96
	New Param.	0.011	0.242	0.97	0.010	0.214	0.98	0.023	0.166	0.97
$\boldsymbol{\sigma} = (0.5, 1, 1), \boldsymbol{\rho} = 0.8$	Univariate	0.006	0.223	0.96	0.023	0.259	0.97	0.038	0.235	0.97
	Original	0.007	0.229	0.97	0.025	0.260	0.97	0.038	0.232	0.97
	New Param.	0.008	0.232	0.97	0.014	0.230	0.97	0.025	0.180	0.97

Table A.8: **Case Study Results with Inverse Wishart Prior**

Model	UI	CIW	SUI	CIW	UUI	CIW
Univariate	0.274 (0.024)	0.096	0.127 (0.022)	0.088	0.066 (0.021)	0.082
Multivariate	0.273 (0.023)	0.090	0.123 (0.019)	0.075	0.058 (0.014)	0.056
New Parameterization	0.273 (0.023)	0.090	0.122 (0.018)	0.069	0.057 (0.014)	0.053

(a) Posterior mean (SD) of marginal prevalences for overall outcome (UI) and two subtypes (SUI, UUI) with corresponding 95% credible interval width (CIW)

	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$
Univariate	0.383 (0.06)	0.435 (0.089)	0.607 (0.131)
Multivariate	0.364 (0.053)	0.401 (0.072)	0.525 (0.093)
	$\hat{\sigma}_1^*$	$\hat{\sigma}_2^*$	$\hat{\sigma}_3^*$
New Param.	0.363 (0.053)	0.641 (0.117)	0.567 (0.111)

	$\hat{\rho}_{12}$	$\hat{\rho}_{13}$	$\hat{\rho}_{23}$
Multivariate	0.462 (0.204)	0.684 (0.157)	0.398 (0.22)
	$\hat{\rho}_{12}^*$	$\hat{\rho}_{13}^*$	$\hat{\rho}_{23}^*$
New Param.	-0.11 (0.251)	0.396 (0.231)	0.093 (0.273)

(b) Posterior mean (SD) of components in estimated covariance matrices for original multivariate model and new parameterization ($\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}}^*$)

Appendix B

Supplementary Materials for "Estimating the Reference Range from a Meta-Analysis"

B.1 Method of moments estimators for lognormal distribution

In (3.9), we use the method of moments estimators for the location and scale parameters of the lognormal distribution in order to transform the observed mean and variance to the log scale, where the observations would be normally distributed. Suppose $Y = \{y_1, ..., y_n\} \sim Lognormal(\mu, \sigma^2)$. Then the first two moments of the lognormal distribution are given by [96]:

$$\begin{aligned} E[Y] &= e^{\mu+\frac{1}{2}\sigma^2} \\ E[Y^2] &= e^{2\mu+2\sigma^2}. \end{aligned} \tag{B.1}$$

We can then set:

$$\begin{aligned} e^{\mu+\frac{1}{2}\sigma^2} &= \frac{1}{n}\sum_{j=1}^{n} y_j = \bar{y} \\ e^{2\mu+2\sigma^2} &= \frac{1}{n}\sum_{j=1}^{n} y_j^2. \end{aligned} \tag{B.2}$$

Solving for μ and σ^2, we have:

$$\begin{aligned} \hat{\mu}_{MM} &= log\left(\frac{\bar{y}^2}{\sqrt{\bar{y}^2 + \frac{n-1}{n}s^2}}\right) = \log\left(\frac{\bar{y}}{\sqrt{1 + \frac{\frac{n-1}{n}s^2}{\bar{y}^2}}}\right) \\ \hat{\sigma}^2_{MM} &= log\left(\frac{\bar{y}^2 + \frac{n-1}{n}s^2}{\bar{y}^2}\right) = log\left(1 + \frac{\frac{n-1}{n}s^2}{\bar{y}^2}\right), \end{aligned} \tag{B.3}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Therefore for each study i in a meta-analyses, we can let:

$$\begin{aligned} \bar{y}_i^* &= \log\left(\frac{\bar{y}_j}{\sqrt{1 + \frac{\frac{n_i-1}{n_i}s_i^2}{\bar{y}_i^2}}}\right). \\ s_i^{2*} &= log\left(1 + \frac{\frac{n_i-1}{n_i}s_i^2}{\bar{y}_i^2}\right) \end{aligned} \tag{B.4}$$

We can then treat $\bar{y}_i^*$ and s_i^{2*} as approximations of the sample mean and sample variance of the study on the log scale.

B.2 Figures

Figure B.1: **WASO Q-Q Plot** Normal Q-Q plot of the study means.

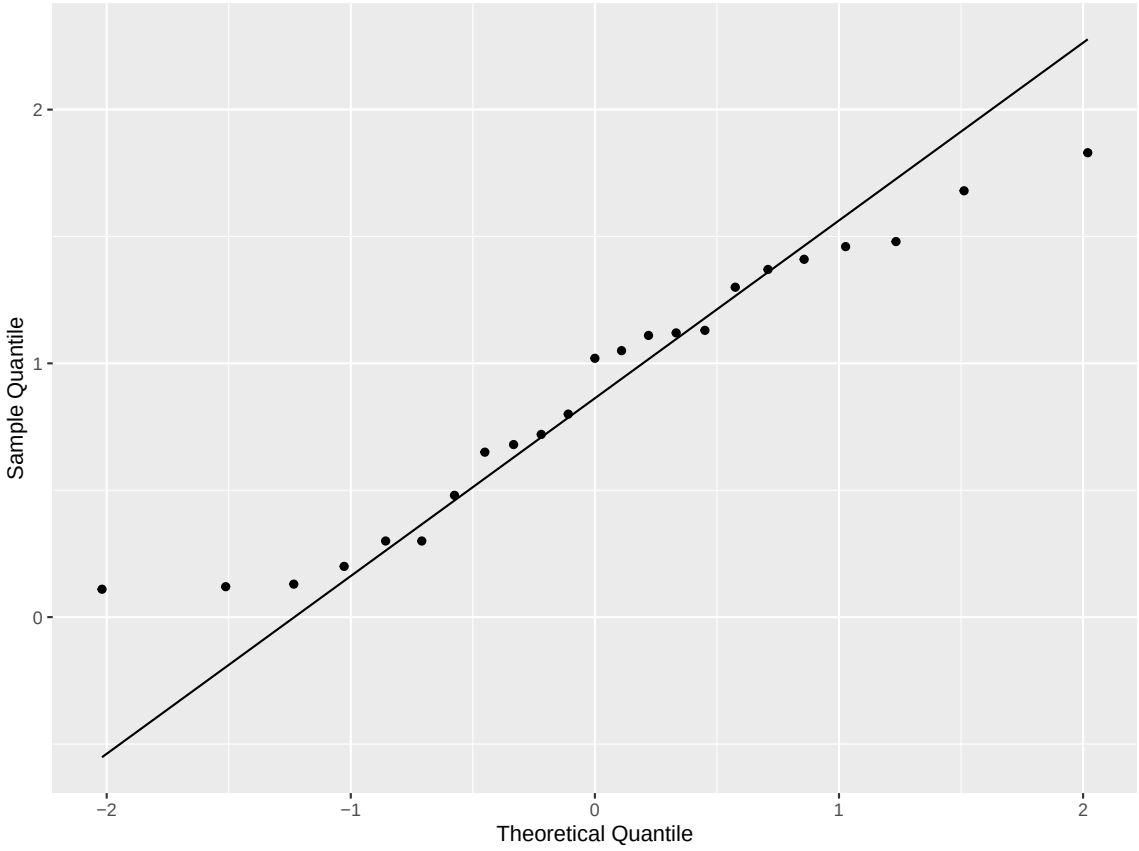
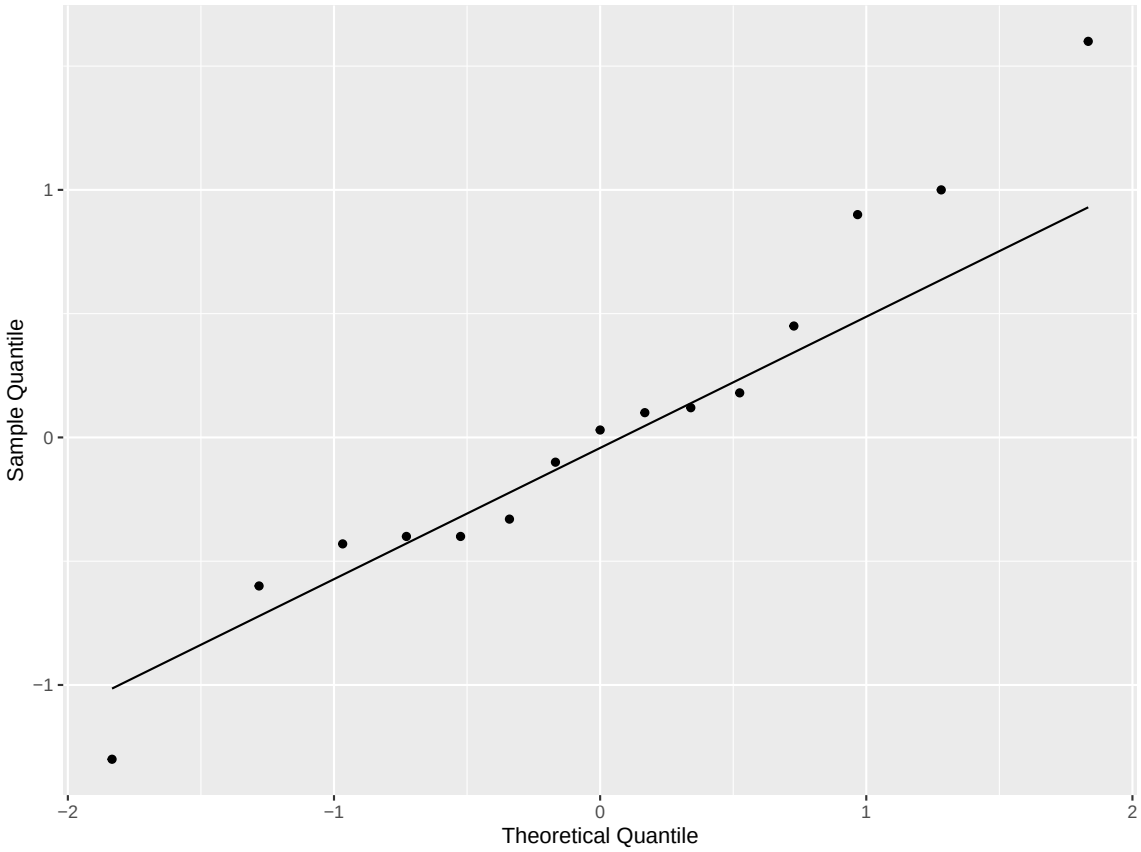


Figure B.2: **SPV Q-Q Plot** Normal Q-Q plot of the study means.



Appendix C

Supplementary Materials for "A Guide to Estimating the Reference Range from a Meta-Analysis using Aggregate or Individual Participant Data"

C.1 Methods for Estimating the Reference Range

C.1.1 Frequentist approach using a random-effects model

Aggregate data

The first method we proposed uses the results of a frequentist random-effects model, which assumes that the underlying study means in the meta-analysis follow a normal distribution with mean μ_{RE} and variance τ^2. We then add the additional assumptions that the data within each study are normally distributed and that the within-study variance (σ^2) is constant across studies. This implies that each of the individuals included in the meta-analysis are marginally distributed $N(\mu_{RE}, \sigma^2 + \tau^2)$. To estimate the reference range, first estimate

the pooled mean across all studies (μ_{RE}) and the between-study variation (τ^2) using a frequentist random-effects model. We use the restricted maximum likelihood (REML) model implemented in the R package "meta" [75], but other methods and software could also be used. Next, use the pooled sample variance as an estimate of the common within-study variance (Table 4.1). Finally, the bounds of an $(1-\alpha)100\%$ level reference range can be estimated by the $\alpha/2$ and $1-\alpha/2$ quantiles of a $N(\mu_{RE}, \sigma^2+\tau^2)$ distribution (Table 4.1).

Individual participant data

If individual participant data are available, any of the methods based on aggregate data described in this guide can still be used by first aggregating the data by study to get the study means, standard deviations, and sample sizes. However, a frequentist random effects model (linear mixed model) can also be fit directly using the individual participant data without first aggregating. We use the R package "lme4," but many other choices of software are available. Let τ^2 be the estimated variance of the random effects, σ^2 be the estimated residual variance, and μ_{RE} be the estimated pooled mean from the model. Then, the bounds of an (1-α)100% level reference range can be estimated by the $\alpha/2$ and 1 - $\alpha/2$ quantiles of a $N(\mu_{RE}, \sigma^2+\tau^2)$ distribution (Table 4.1).

C.1.2 Bayesian posterior predictive interval

Aggregate data

The second method we proposed uses the posterior predictive distribution of a new individual from a Bayesian random-effects model. This imposes the same distributional assumptions as with the frequentist approach. The sampling distributions of the study-means and standard deviations can be used to estimate the posterior distributions of μ_{RE}, σ^2, and τ^2 using Markov Chain Monte Carlo sampling (Table 4.1). An (1-α)100% level reference range can then be estimated by the $\alpha/2$ and 1 - $\alpha/2$ quantiles of samples from a $N(\mu_{RE}, \sigma^2+\tau^2)$ distribution, the posterior predictive distribution for a new individual. We place a N(0,

1000) prior on μ_{RE}, and Uniform(0,100) priors on σ and τ, as shown in Table 4.1. The main difference between this and the frequentist methods is that the posterior predictive interval incorporates the uncertainty in the estimated parameters into the reference range, whereas the frequentist methods do not.

Individual participant data

However, if individual participant data are available, a Bayesian random effects model can also be fit directly on the individual observations, just as with the frequentist approach. Instead of using the sampling distributions for the study means and standard deviations, we can use the likelihood for an individual observation (Table 4.1). We still place the same priors on each of the estimated parameters, and the resulting range has the same interpretation as the posterior predictive interval based on the aggregate data.

C.1.3 Empirical approach

Aggregate data

Finally, we proposed a simple empirical approach using aggregate data, which is similar to the method used by Conceição et al. [9] to estimate reference ranges for normal Subjective Postural Vertical (SPV) measurements. This does not make the same assumption about constant within-study variance, but still assumes the data are normally distributed. First, empirically estimate the pooled mean (μ_{emp}), weighting by the sample size in each study. This is equivalent to the mean estimate in the fixed effects model proposed by Laird and Mosteller [64] when weighting by sample size. Then, estimate the total variance both within and across studies (σ_T^2) (Table 4.1). An (1-α)100% level reference range can then be estimated by the $\alpha/2$ and 1 - $\alpha/2$ quantiles of a $N(\mu_{emp}, \sigma_T^2)$ distribution.

Individual participant data

If individual participant data are available, one could equivalently pool the data across studies and estimate the pooled mean (μ_{emp}) as the mean of these individual measurements. The total variance within and across studies could similarly be estimated as the variance of these pooled samples (σ_T^2) (Table 4.1). Then, an (1-α)100% level reference range can be estimated by the $\alpha/2$ and 1 - $\alpha/2$ quantiles of a $N(\mu_{emp}, \sigma_T^2)$ distribution.

C.2 Lognormally distributed data

In some cases, such as when a measurement cannot take on negative values, it may be more reasonable to assume that the data within each study follow a lognormal distribution. If individual participant data are available, the preferred approach would be to first log-transform the individual observations, estimate the reference interval, then exponentiate the resulting bounds. However, if only aggregate data are available, the observed means and standard deviations need to be transformed to the log scale. Suppose $Y = \{y_1, \ldots, y_n\}$ denotes a set of continuous observations. Because, $\frac{1}{n}\sum_{i=1}^{n} log(y_i) \neq log\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$, the observed means and sample variances on the log-scale must be estimated. The method of moments estimators for the mean and variance of log(Y) are given in Table 4.3. These equations can be used to estimate the observed means and sample variances, which can then be used in each of the methods described to estimate the reference range. Finally, the resulting range can be exponentiated in order to return to the original scale. We note that when performing either of these transformations, the normality assumption for the study means now applies to the log-transformed data and should be still be assessed using methods such as a normal Q-Q plot, as the transformed means may be skewed.

Table C.1: **Estimating the Study Means and Sample Variances on the Log Scale with Aggregate Data**

Let $\overline{y_i}$, s_i^2 , and n_i be the sample mean, sample variance, and sample size for study *i*, respectively. The method of moments estimators for the location and scale parameters of the log-normal distribution are given by*:*

$$\overline{y_i}^* = \log\left(\frac{\overline{y_i}}{\sqrt{1 + \frac{\frac{n_i - 1}{n_i} s_i^2}{\overline{y_i}^2}}}\right)$$

$$s_i^{2^*} = \log\left(1 + \frac{\frac{n_i - 1}{n_i} s_i^2}{\overline{y_i}^2}\right)$$

We can then use $\overline{y_i}^*$ and $s_i^{2^*}$ as estimates of the mean and sample standard deviation on the log scale when individual participant data are not available.

Note: We assume $s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$

C.3 Clinical Scenario

C.3.1 Tables

Table C.2: **Aggregate Data for Liver Stiffness Example**

	Study	Mean	SD	Sample Size
1	1.00	4.66	1.38	581
2	2.00	4.11	0.89	530
3	3.00	3.97	1.00	67
4	4.00	4.57	1.29	248
5	5.00	4.53	1.30	206
6	6.00	4.99	1.08	183
7	7.00	4.20	1.18	60
8	8.00	4.67	1.18	35
9	9.00	5.05	4.42	34
10	10.00	4.82	1.22	132
11	11.00	5.20	1.39	420
12	12.00	4.83	1.25	90
13	13.00	5.09	1.18	433
14	14.00	4.52	1.51	498
15	15.00	5.45	1.87	52
16	16.00	5.18	0.68	29
17	17.00	5.17	1.13	9
18	18.00	3.83	0.67	15
19	19.00	4.36	1.37	26

Table C.3: **Sensitivity Analysis with Aggregate Data** Results when removing studies 9 and 16 and estimating reference ranges using aggregate data.

	95% Reference Range
Frequentist AD	(2.58, 7.74)
Bayesian AD	(2.57, 7.84)
Empirical AD	(2.61, 7.75)

Table C.4: **Sensitivity Analysis with IPD** Results when removing studies 9 and 16 and estimating reference ranges using IPD.

	95% Reference Range
Frequentist AD	(2.62, 7.74)
Bayesian AD	(2.61, 7.79)
Empirical AD	(2.64, 7.69)
Frequentist IPD	(2.63, 7.72)
Bayesian IPD	(2.52, 7.94)
Empirical IPD	(2.64, 7.69)

C.3.2 Figures

Figure C.1: **Forest Plot of Study Standard Deviations from Clinical Scenario** Standard deviations of the log of liver stiffness measurements for each study and corresponding 95% confidence intervals. The observed standard deviations in studies 9 and 16 look as though they may differ from the others. The vertical dotted line represents the estimated pooled standard deviation.

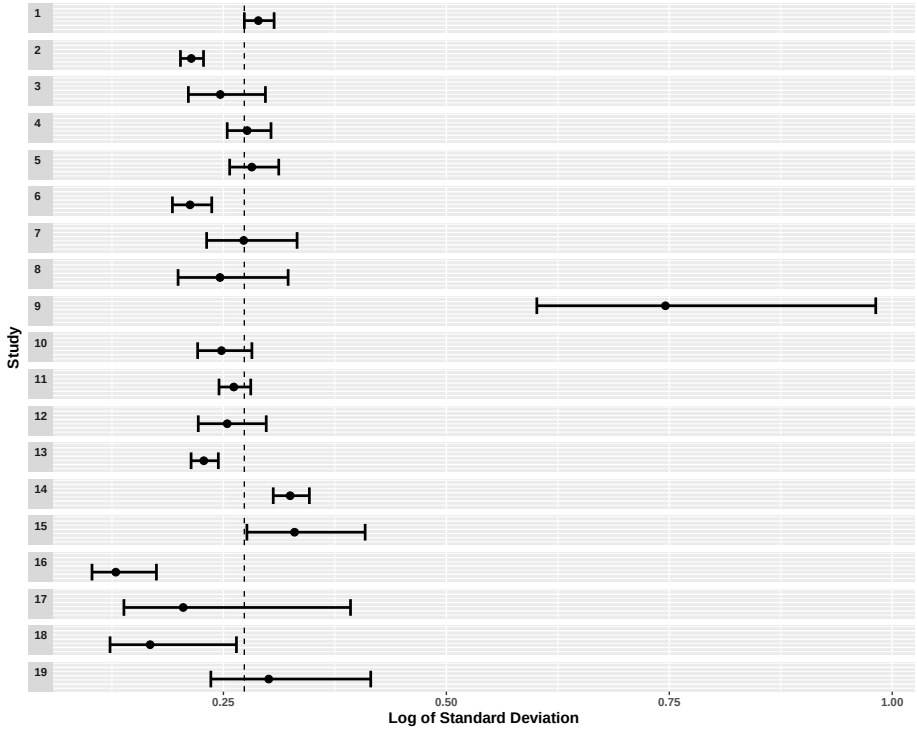


Figure C.2: **Normal Q-Q plot of log-transformed means of liver stiffness**

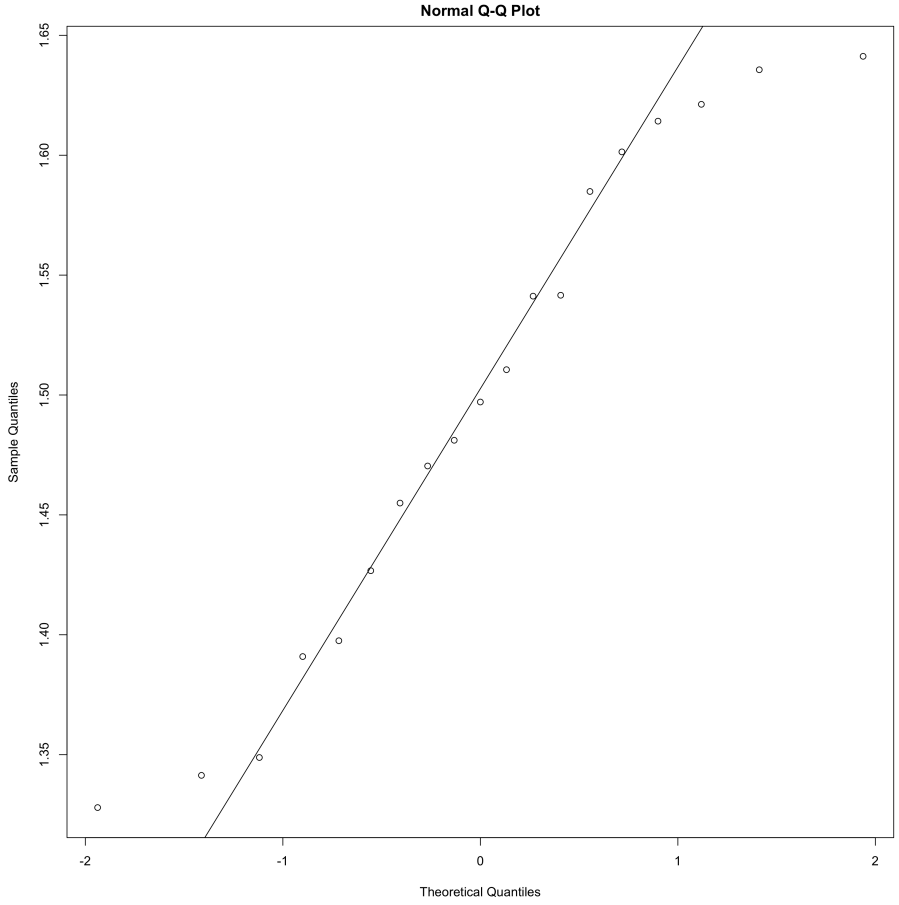


Figure C.3: **Histogram of the pooled log-transformed liver stiffness measurements across all 19 studies.**

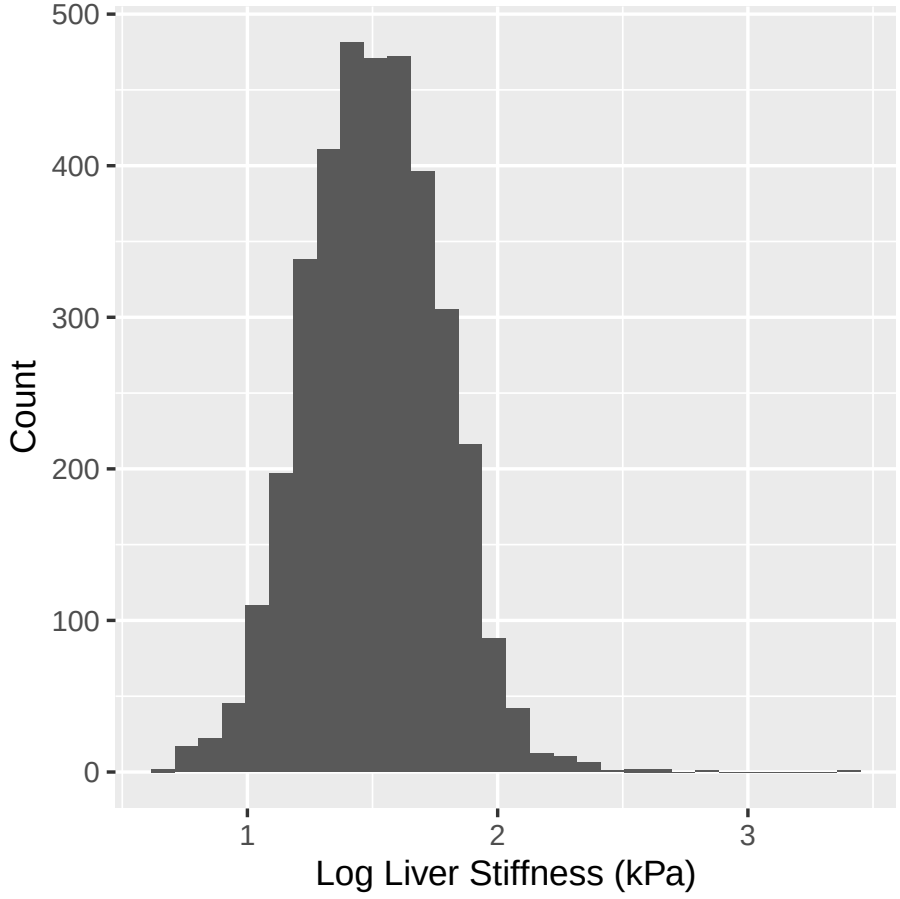


Figure C.4: **Histograms of the log-transformed liver stiffness measurements by study.**