The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2993–3012
https://doi.org/10.1214/23-AOAS1748
© Institute of Mathematical Statistics, 2023

BUILDING A DOSE TOXO-EQUIVALENCE MODEL FROM A BAYESIAN META-ANALYSIS OF PUBLISHED CLINICAL TRIALS

By Elizabeth A. Sigworth[1,a], Samuel M. Rubinstein[2,c], Jeremy L. Warner[3,d], Yong Chen[4,e] and Qingxia Chen[1,b]

[1]*Department of Biostatistics, Vanderbilt University,* [a]*elizabeth.a.sigworth@vanderbilt.edu,* [b]*cindy.chen@vumc.org*
[2]*Division of Hematology, University of North Carolina School of Medicine,* [c]*samuel_rubinstein@med.unc.edu*
[3]*Department of Medicine, Vanderbilt University School of Medicine,* [d]*jeremy.warner@vumc.org*
[4]*Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania,* [e]*ychen123@pennmedicine.upenn.edu*

In clinical practice medications are often interchanged in treatment protocols when a patient negatively reacts to their first line of therapy. Although switching between medications is common, clinicians often lack structured guidance when choosing the initial dose and frequency of a new medication, given the former with respect to risk of adverse events. In this paper we propose to establish this dose toxo-equivalence relationship using published clinical trial results with one or both drugs of interest via a Bayesian meta-analysis model that accounts for both within- and between-study variances. With the posterior parameter samples from this model, we compute median and 95% credible intervals for equivalent dose pairs of the two drugs that are predicted to produce equal rates of an adverse outcome, relying solely on study-level information. Via extensive simulations, we show that this approach approximates well the true dose toxo-equivalence relationship, considering different study designs, levels of between-study variance, and the inclusion/exclusion of nonconfounder/nonmodifier subject-level covariates in addition to study-level covariates. We compare the performance of this study-level meta-analysis estimate to the equivalent individual patient data meta-analysis model and find comparable bias and minimal efficiency loss in the study-level coefficients used in the dose toxo-equivalence relationship. Finally, we present the findings of our dose toxo-equivalence model applied to two chemotherapy drugs, based on data from 169 published clinical trials.

1. Introduction.

Often, there are multiple medication regimens that can be prescribed to patients to treat the same type of illness. However, these regimens can differ in their dosing as well as in their risks of inducing adverse events in patients. As a motivating example, we concern ourselves with the taxane chemotherapy drugs paclitaxel and docetaxel, which are both known to induce peripheral sensory neuropathy, an outcome that is believed to be directly related to cumulative exposure. Patients are frequently switched from paclitaxel to docetaxel, due to infusion reactions, yet there currently exists no clear guidance on how clinicians should choose an initial dosage and frequency of docetaxel, given a patient's previous paclitaxel regimen. However, as with the side effects for many drugs in similar scenarios, the incidence rate of peripheral sensory neuropathy in clinical trials of paclitaxel and docetaxel is commonly reported in published literature. Thus, it is desired to develop a method that leverages this available pool of study meta-information to estimate the dose toxo-equivalence relationship.

Conventional meta-analysis approaches combine results from independent studies to find patterns or discrepancies in the published literature (Hedges and Olkin (1985)). They typi-

Received July 2021; revised January 2023.
Key words and phrases. Bayesian methods, dose toxo-equivalence, individual patient data meta-analysis, study-level meta-analysis.

cally use summary statistics reported by each study, such as effect size estimates and standard errors, in either common effects or random effects models to make inferences on the true value of the parameter of interest (Hedges and Olkin (1985)). In the case of clinical trials, the reported summary statistics are often treatment effect estimates, such as odds ratios or risk differences, between groups exposed to the two drugs of interest within the same study (Whitehead (2002)). As we are interested in the dose relationship and not the treatment effect, we need to extract the response rate for the treatment and its associated dosage information from each study rather than estimated treatment effect. Ultimately, we need a method to incorporate reported incidence rates for each drug at their specific dosage as well as potentially aggregated summary data from any study in one or both of these drugs in a way that adds little bias, loses minimal efficiency, and produces a useful approximation of this dose toxo-equivalence.

The use of aggregated summary data in study-level meta-analysis has the potential to induce bias. Focused on treatment effect estimates, Berlin et al. (2002) investigated a real-world published example for which both individual patient-level and study-level data were available and ultimately recommended using individual patient data, when feasible, to study patient characteristics to avoid aggregation bias (in the presence of effect modifiers). In this paper we will use the term individual patient data (IPD) meta-analysis for the regression analysis based on individual patients and refer to the analysis based on aggregated study-level data as study-level (SL) meta-analysis. When there is no interaction effect between patient characteristics and treatment, another question researchers asked was when IPD meta-analysis and SL meta-analysis would yield identical results. Among others, Olkin and Sampson (1998) and Steinberg et al. (1997) illustrated some special settings for which IPD meta-analysis and SL meta-analysis could generate identical treatment effect estimators, all assuming continuous outcomes in the IPD meta-analysis. When all covariates are at the study-level, SL and IPD analyses are generally equivalent. However, when there are individual-level factors or when these factors are summarized at the individual level, IPD meta-analysis is preferred, in practice, over SL meta-analysis, due to the risk of aggregation bias, particularly in the presence of an effect modifier (Berlin et al. (2002), Lambert et al. (2002)). While IPD meta-analysis is preferred in these scenarios, oftentimes the IPD from different studies are difficult to obtain. Additionally, the size of the data used in IPD meta-analysis can lead to high computation time, compared to SL meta-analysis, particularly with Bayesian posterior sampling. For these reasons SL meta-analysis is more practical and feasible, especially when there exists no prior evidence of effect modifiers.

There is increasing interest in studying the relative efficiency of meta-analyses, based on fitted results at the study level compared to IPD meta-analysis, given barriers on data sharing and/or protections on participant privacy. Among others, Lin and Zeng (2010b) showed that, when compared to IPD meta-analysis approaches (called mega-analysis in their setting), common effect meta-analysis methods using effect estimates from models fit at the study-level have minimal efficiency loss, and Zeng and Lin (2015) further showed that random effect meta-analysis methods are at least as efficient as the former. However, their results do not apply to our setting. In particular, Lin and Zeng (2010b) considered a regression (see their first equation in Section 2.1) of an outcome Y_{ki} on covariates X_{ki}, for the ith participant in the kth study. Implicitly, they assumed covariates collected within each study (X_{ki}) had variations within kth study, whereas in our setting, within a study, the dosage is fixed for the corresponding treatment (i.e., there is no variation within kth study).

While various meta-analysis approaches have been intensively studied, little work has been done in the dose-equivalence model setting. The validity of SL meta-analysis under this setting has not been studied and its relative efficiency vs. IPD meta-analysis has not been evaluated. We aim to fill this gap by developing a Bayesian random-effects model to study the dose toxo-equivalence relationship.

In this paper we distinguish the covariates with no variation at the subject level (such as treatment and designed dosage) in the aggregated study-level data from those with variations (such as percentage of male or mean age) and call the former study-level covariates and the latter subject-level covariates. In Section 3 we propose a Bayesian random-effects SL meta-analysis model that accounts for both within- and between-study variances, with or without additional subject-level covariates. We show that under the proposed model the toxo-equivalence curve depends solely on the coefficients of the study-level covariates. Based on the posterior samples produced by this model, we compute median and 95% credible intervals for equivalent dose pairs of any two drugs of interest that are predicted to result in equal rates of the adverse outcome. Via extensive simulation studies in Section 4, we demonstrate the ability of this model to closely approximate the true dose toxo-equivalence relationship for different study designs, varying levels of between-study variance, and in the presence of subject-level data (which are not treatment or dose effect modifiers) in addition to study-level covariates. We compare the performance of this meta-analysis approach in terms of bias and efficiency to an IPD meta-analysis model fit on pooled subject-level information, and demonstrate that our method results in comparable levels of bias to the IPD approach, as well as minimal efficiency loss in all parameters used to calculate the dose toxo-equivalence relationship when the model is correctly specified. Additionally, we consider the sensitivity of our meta-analysis approach to various types of model misspecification. Finally, we illustrate our method with empirical data gathered from published clinical trials in either paclitaxel or docetaxel in Section 5. We conclude the paper with discussion in Section 6.

2. Motivating example.

Our motivating example looks at the chemotherapy medications paclitaxel and docetaxel, both members of the taxane class of drugs that are prescribed to treat a variety of cancers (Warner et al. (2015)). Taxanes are known to induce peripheral sensory neuropathy, with patient risk for this outcome believed to be directly related to cumulative exposure (Argyriou et al. (2008)). Clinicians frequently start patients on paclitaxel as their first line of therapy, but some are unable to continue treatment, due to hypersensitivity or infusion reactions, at which point patients are often switched to a docetaxel treatment regimen. However, there is no clear guidance on how to choose the initial dose and schedule of docetaxel, given a previous paclitaxel regimen, particularly with respect to the overall risk of peripheral sensory neuropathy. Since the rate of neuropathy development within studies is commonly reported as an adverse effect of treatment, we performed a systematic review of randomized or nonrandomized clinical trials of paclitaxel or docetaxel monotherapy among cancer patients aged ≥ 18 years, extracting all aggregated data necessary for the dose toxo-equivalence calculation. Individual patient data from the included studies was not attainable. We apply the method for study-level data described in Section 3 to this data in Section 5, after exploring its performance compared to individual patient data. Further insight into the clinical relevance of our approach, including complete details of the systematic review procedure and examples of our method applied to specific chemotherapy regimens, can be found in our related clinical paper (Sigworth et al. (2022)).

3. Methods.

3.1. *Hierarchical model structure.*

3.1.1. *IPD meta-analysis.* We first consider the subject-level IPD meta model against which we will compare our SL meta approach. Denote $D_{ij} = (X_{iA}, X_{iB}, d_i, \mathbf{Z_{ij}})$, $i = 1, \dots, N$, $j = 1, \dots, n_i$ as the subject-level covariates, where X_{iA} is an indicator that study i uses drug A, X_{iB} is an indicator that study i uses drug B, d_i is the dose received in study

i normalized to mean 0 and standard deviation 1 (or a normalized version of a monotone transformation of the dose such as the square-root transformation), and $\mathbf{Z_{ij}}$ is a vector of subject-specific potential covariates assumed to be associated with the adverse event. Dose values were normalized to improve computational efficiency in the Bayesian fitting procedure (Kruschke (2015)). Additionally, N is the total number of studies, and n_i is the number of subjects in study i. Let w_{ij} denote the incidence indicator of the adverse event of interest, with $w_{ij} = 1$ for a subject experiencing the adverse event and $w_{ij} = 0$ otherwise. We assume $w_{ij}|p_{ij} \sim \text{Bernoulli}(p_{ij})$ where without additional covariates we have

$$\text{logit}(p_{ij}) = \mu_i + \alpha_1 + \alpha_2 X_{iB} + \alpha_3 X_{iA} d_i + \alpha_4 X_{iB} d_i,$$

and with additional covariates we have

$$\text{logit}(p_{ij}) = \mu_i + \alpha_1 + \alpha_2 X_{iB} + \alpha_3 X_{iA} d_i + \alpha_4 X_{iB} d_i + \boldsymbol{\alpha}_z' \mathbf{Z_{ij}}.$$

In this model α_1 is the mean outcome for studies in drug A with a normalized dose of 0, and $\alpha_1 + \alpha_2$ is the mean outcome for studies in drug B with a normalized dose of 0. We also estimate a random intercept component for each study, μ_i, as a measure of between-study heterogeneity. Its variance, τ^2, measures between-study variance in responses not attributable to other included variables. Note that throughout this manuscript we use the subscript i to denote the study of interest, that is, the group of subjects assigned to the same protocol within the same study. In this way single-arm and multiarm studies can be analyzed via this method. Noninformative priors of $\mu_i|\tau \sim N(0, \tau^2)$, with $\tau \sim \text{InvGamma}\ (0.001, 0.001)$ and $\boldsymbol{\alpha} \sim \text{MVN}(\mathbf{0}, 10^6\, \text{diag}(\mathbf{1}))$, are specified, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_4, \boldsymbol{\alpha}_z')$. Under this model our posterior distribution for $\boldsymbol{\alpha}, \tau$ is

$$p(\boldsymbol{\alpha}, \tau|\mathbf{D_{ij}}, w_{ij}) \propto p(w_{ij}|\mathbf{D_{ij}}, \mu_i, \boldsymbol{\alpha}) p(\mu_i|\tau) p(\tau) p(\boldsymbol{\alpha}).$$

3.1.2. *Study-level meta-analysis.* Next, we propose a Bayesian hierarchical model of the prevalence of an adverse event across multiple studies using aggregated meta-data. Let Π_i represent the rate of the adverse event in study i, $\Pi_i = \frac{1}{n_i} \sum_{j=1}^{n_i} w_{ij}$, and define $Y_i = \text{logit}(\Pi_i)$. Denote the per-study data as $\mathbf{D_i} = (X_{iA}, X_{iB}, d_i, \mathbf{Z_i^a})$, where X_{iA}, X_{iB}, and d_i are as defined previously and are study-level variables taking the same value for all subjects in study i, and $\mathbf{Z_i^a}$ is a vector of aggregated summary statistics of $\mathbf{Z_{ij}}$ at each study i, such as frequencies of categorical variables or means of continuous variables. Then the outcome Y_i can be modeled as $Y_i|\mu_i, \boldsymbol{\beta}, \mathbf{D_i} \sim N(\phi_i, S_i^2)$, where

$$\phi_i = \mu_i + \beta_1 + \beta_2 X_{iB} + \beta_3 X_{iA} d_i + \beta_4 X_{iB} d_i + \boldsymbol{\beta}_Z' \mathbf{Z_i^a}.$$

Here S_i^2 is the within-study variance of our outcome, which depends on the size n_i of the relevant arm of the study as well as the count of adverse events in that outcome, k_i, such that $S_i^2 = 1/k_i + 1/(n_i - k_i)$ (the variance of a logit transformed proportion). As previously, β_1 represents the mean outcome for studies in drug A with a normalized dose of 0 and $\beta_1 + \beta_2$ is the mean outcome for studies in drug B with a normalized dose of 0, while μ_i represents a random intercept component estimated at the study level to measure between-study heterogeneity. We set noninformative priors of $\mu_i|\tau \sim N(0, \tau^2)$, $\tau \sim \text{InvGamma}(0.001, 0.001)$, and $\boldsymbol{\beta} \sim \text{MVN}(\mathbf{0}, 10^6\, \text{diag}(\mathbf{1}))$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_4, \boldsymbol{\beta}_Z')$. Note that although noninformative priors were used in our simulations (and case study), informative priors could be considered if prior knowledge on the dose toxo-equivalence relationship was available. Based on this structure, the posterior distribution for $\boldsymbol{\beta}, \tau$ is

$$p(\boldsymbol{\beta}, \tau|\mathbf{D_i}, Y_i) \propto p(Y_i|\mathbf{D_i}, \mu_i, \boldsymbol{\beta}) p(\mu_i|\tau) p(\tau) p(\boldsymbol{\beta}).$$

When including additional covariates in the data-generating model, we consider SL meta-analysis models both with and without $\mathbf{Z_i^a}$, denoted SL-C and SL-NC, respectively. The aggregated study covariates in the SL-C model are intended to adjust for the additional heterogeneity in responses that may be due to these values but not to estimate $\boldsymbol{\alpha}$ from the IPD meta-analysis. By considering both SL-C and SL-NC in the presence of additional covariates $\mathbf{Z_i^a}$, we can evaluate the sensitivity of our method to this extra information.

3.2. *Equivalence relationship.* For our IPD meta-analysis fits, the hierarchical structure produces expected outcomes of

$$\begin{aligned}&\operatorname{logit}[E(w_{ij}|X_{iA}=1, X_{iB}=0, d_i=d_A, \mathbf{Z_{ij}}=\mathbf{z})] = \mu+\alpha_1+\alpha_3 d_A+\boldsymbol{\alpha}_z'\mathbf{z},\\ &\operatorname{logit}[E(w_{ij}|X_{iA}=0, X_{iB}=1, d_i=d_B, \mathbf{Z_{ij}}=\mathbf{z})] = \mu+\alpha_1+\alpha_2+\alpha_4 d_B+\boldsymbol{\alpha}_z'\mathbf{z}\end{aligned}$$

for studies in drugs A and B, respectively, where the study in drug A had dose d_A and the study in drug B had dose d_B. From these expected outcomes, we can build a dose toxo-equivalence model for the dose and adverse outcome relationship between the two drugs of interest. As the logit is a monotone transformation, we look to solve

$$\begin{aligned}&\operatorname{logit}[E(w_{ij}|X_{iA}=1, X_{iB}=0, d_i=d_A, \mathbf{Z_{ij}}=\mathbf{z})]\\ &\quad= \operatorname{logit}[E(w_{ij}|X_{iA}=0, X_{iB}=1, d_i=d_B, \mathbf{Z_{ij}}=\mathbf{z})]\end{aligned}$$

with respect to d_B, assuming that d_A is known, which, when simplified, results in solving

$$d_{B,IPD} = (\alpha_3 d_A - \alpha_2)/(\alpha_4).$$

As for the SL meta-analysis, from our aggregated meta fits we can produce the expected outcomes for studies in drugs A and B as

$$\begin{aligned}&E(Y_i|X_{iA}=1, X_{iB}=0, d_i=d_A, \mathbf{Z_i^a}=\mathbf{z}) = \mu+\beta_1+\beta_3 d_A+\boldsymbol{\beta}_Z'\mathbf{z},\\ &E(Y_i|X_{iA}=0, X_{iB}=1, d_i=d_B, \mathbf{Z_i^a}=\mathbf{z}) = \mu+\beta_1+\beta_2+\beta_4 d_B+\boldsymbol{\beta}_Z'\mathbf{z},\end{aligned} \tag{1}$$

where the doses in the studies are d_A and d_B, respectively, as before. We look to solve $E(Y_i|X_{iA}=1, X_{iB}=0, d_i=d_A, \mathbf{Z_i^a}=\mathbf{z}) = E(Y_i|X_{iA}=0, X_{iB}=1, d_i=d_B, \mathbf{Z_i^a}=\mathbf{z})$ with respect to d_B, which results in

$$d_{B,SL} = (\beta_3 d_A - \beta_2)/(\beta_4). \tag{2}$$

For each fit we vary across a range of plausible values of d_A to find dose pairs (d_A, d_B) for which we expect the rates of adverse outcome to be equivalent, generating a posterior distribution for d_B. We use the Markov chain Monte Carlo methods, employed by the JAGS software package in R, to produce posterior samples for each of our model parameters. Let β_i^k be the kth MCMC sample from the posterior, then $d_{B,SL}^k = (\beta_3^k d_A - \beta_2^k)/(\beta_4^k)$.

We report the median and 95% credible intervals from our calculated distributions of $\{d_B^k, k=1,\dots,K\}$ with K being the total number of draws from the posterior. Similarly, we calculate dose pairs $(d_A, d_{B,IPD}^k)$ from the posterior of our IPD meta-analysis.

Note that under the additive model assumption in (1), the dose toxo-equivalence relationship in (2) is independent of the common intercept β_1 and the coefficients of the aggregated covariates, $\boldsymbol{\beta}_Z$. This is critical because, as we will demonstrate in Section 4, there is minimal efficiency loss or increase in bias in estimating $(\beta_2, \beta_3, \beta_4)$, using SL meta-analysis when comparing to IPD meta-analysis, resulting in comparable dose toxo-equivalence relationship curves with the correctly specified model.

4. Simulation study.

4.1. *Simulation for a correctly specified model.*

4.1.1. *Method.* To evaluate the performance of our method in terms of bias, efficiency, and ability to recover the true dose toxo-equivalence relationship, we performed extensive simulations, varying study types (balanced and unbalanced), generating models (with or without subject-level covariates), and levels of between-study variability ($\tau^2 \in \{0, .5, 1\}$). We performed 500 repetitions for each combination, for a total of 6000 simulations.

For each setting we first simulated data from $N = 150$ study datasets, 75 in each drug. For consistency we sampled 150 dose values from a $U(-1, 1)$ distribution, which are used for each simulation repetition. Next, a random intercept $\mu_i, i \in \{1, \dots, 150\}$ is generated for each of the 150 studies from a $N(0, \tau^2)$ distribution. If the setting requires balanced studies, each of the 150 study-level datasets will contain $n_i = 100$ subjects; if unbalanced, each study contains between 50 and 200 subjects, in increments of 10, with equal probability. Each subject in study i is assigned the same X_{iA} and X_{iB} indicator variables, based on the drug type under study in study i, that is, if study i is in drug A, then $X_{iA} = 1$ and $X_{iB} = 0$. If the study includes additional subject-level covariates, then each subject j has two subject-specific independent covariates drawn: a binary Z_{ij5} with study-specific probability θ_i drawn from Unif(0.2, 0.5) and a continuous Z_{ij6} drawn from $N(0, 1)$.

Once the subject-level covariate data is generated, we calculate subject-specific log-odds of the adverse event for subject j in study i with normalized dose d_i, based on a preselected vector of study-level coefficients $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (-0.6, -0.8, -0.5, -0.9)$, and subject-level coefficients $(\alpha_5, \alpha_6) = (0.2, 0.5)$, the latter to be included in studies with additional covariates. These parameters were based on those estimated in our clinical application, discussed in Section 5, and produce nonrare event outcomes. Thus, for a generating model without additional covariates, the subject-specific log-odds are

$$\operatorname{logit}(p_{ij}) = \mu_i - 0.6 - 0.8X_{iB} - 0.5X_{iA}d_i - 0.9X_{iB}d_i,$$

and for a generating model with two additional subject-level covariates, we have

$$\operatorname{logit}(p_{ij}) = \mu_i - 0.6 - 0.8X_{iB} - 0.5X_{iA}d_i - 0.9X_{iB}d_i + 0.2Z_{ij5} + 0.5Z_{ij6}.$$

From $\operatorname{logit}(p_{ij})$ we use the expit function to produce subject-specific probabilities of the adverse outcome, p_{ij}. We then generate adverse outcome indicators, w_{ij}, from a Bernoulli$(1, p_{ij})$. From this simulated dataset at the subject-level, we summarize to reflect the metrics commonly reported in a clinical trial. Dose d_i is the same across subjects within a study and does not need to be summarized. For binary covariate Z_{ij5}, we take the mean across all subjects in study i, and for continuous covariate Z_{ij6}, we take the median, creating study-level summary values Z^a_{i5} and Z^a_{i6}. Event rate per study is calculated as $\Pi_i = \frac{1}{n_i}\sum_{j=1}^{n_i} w_{ij}$. Studies with $\Pi_i \in \{0, 1\}$ are dropped and a new dataset generated until 75 valid datasets in each drug have been created (fewer than one in 500 simulated studies were regenerated). Finally, the logit of the outcome, $\operatorname{logit}(\Pi_i)$, is calculated, along with $\operatorname{Var}(\operatorname{logit}(\Pi_i)) = S_i^2 = 1/k_i + 1/(n_i - k_i)$, where $k_i = \sum_{j=1}^{n_i} w_{ij}$.

Next, for the generating model without additional covariates, we fit our Bayesian models on both the aggregated study-level meta-data and the subject-level IPD meta-data. All models were fit using JAGS version 4.3.0 in R version 3.6.0 with packages R2jags version 0.6-1 and coda version 0.19-2 and were executed using the Vanderbilt Advanced Computing Center for Research and Education (ACCRE) cluster. Each model consists of four independent chains. Where $\tau^2 = 0$, we use a burn-in of 5000 and take 20,000 samples with a thinning interval of 2, based on within-chain correlation from preliminary fits. For $\tau^2 \in (0.5, 1)$, we use a burn-in of 10,000 and take 40,000 samples with a thinning interval of 4. All R code necessary to reproduce these simulations can be found at https://github.com/esigworth/BayesianDTEM as well as in the Supplementary Material (Sigworth et al. (2023)).

For the generating model with additional covariates, we fit three candidate models: an SL-C meta-analysis model including aggregated Z^a_{i5} and Z^a_{i6}, an SL-NC meta-analysis model fit on only study-level covariates, and the subject-level IPD meta-data model incorporating study-level covariates and subject-level Z_{ij5} and Z_{ij6}. Each of these models was fit using the same burn-in, sampling, and thinning settings as above, based on τ^2.

From the full set of posterior samples of each parameter in each model, we calculate the dose toxo-equivalence relationship across a range of normalized doses between -1 and 1, saving the median 95% credible interval bounds from each simulation. For model diagnostics we report coverage probabilities, the ratio of the median absolute deviation (MADR) for each parameter in each SL meta fit to the MAD of the IPD meta fit, mean square error, percent bias, and relative efficiency. Percent bias is calculated as the median across 500 simulations of the difference between the mean of the posterior distribution for a given parameter and the true value of that parameter, divided by the true value and multiplied by 100. Relative efficiency is the median across simulations of the ratio of the variance in the posterior of the IPD model to the variance in the SL model (equivalent to the efficiency of the SL model over the efficiency of the IPD model, where efficiency is the inverse of variance). Both percent bias and relative efficiency are reported with 95% credible intervals.

4.1.2. *Results.* Our summary of the performance of this method focuses on the parameters involved in the calculation of the dose toxo-equivalence relationship, $(\beta_2, \beta_3, \beta_4)$ and $(\alpha_2, \alpha_3, \alpha_4)$, since it is only the performance of these parameters that will determine the ability of our meta and IPD meta models to approximate the true dose toxo-equivalence relationship.

We first assess our method in the absence of additional covariates. Table 1 summarizes the coverage probabilities and the MADRs and 95% credible interval widths (MADR CIW) for each parameter in each setting, arranged by study design and value of τ^2. Coverage is near 95% for all models and settings, and the MADR values for all fits are close to 1 with narrow MADR CIWs, signifying comparable levels of variability within the posterior sampling chains of our two model approaches.

In Figure 1(A) we display the median and 95% credible intervals for percent bias in parameter estimates from the SL meta and IPD meta model fits without additional covariates, with three levels of τ^2 $(0, 0.5, 1)$ displayed across the columns and study designs across the rows. Median percent bias is consistently close to zero, with variance around the median increasing with increasing τ^2. Figure 1(B) presents the median and 95% credible intervals for the relative efficiency of the SL meta model to the IPD meta model with no additional covariates. Efficiency is consistently close to 1, shifting slightly higher with increasing τ^2 but with credible intervals always containing 1, indicating comparable efficiency.

In Figure 1(C) we present the estimated dose toxo-equivalence curves and 95% credible intervals for the IPD meta (pink) and SL meta (blue) model fits with no additional covariates and compare these to the true dose toxo-equivalence relationship (green), by τ^2 (columns) and study design (rows). Credible interval widths increase with τ^2 in all settings. The SL and IPD estimated relationship curves are very similar to both the true model and one another across settings. With no additional covariates, then the performance of the IPD meta and SL meta models in recovering the true dose toxo-equivalence relationship is comparable.

Next, we look at the performance of our method when additional subject-level covariates are included in the generating model, comparing the IPD meta fit on complete data to the SL-NC and SL-C models on study-level and potentially summarized data. Looking at

TABLE 1

Coverage, median absolute deviation ratios (MADR), percent bias (IQR), and MSE, by fit type and study design, for simulations without additional covariates. Target coverage was 0.95

Simulations Without Additional Covariates								
	Statistic	Coverage		MADR	Percent Bias (IQR)		MSE	
Setting	Fit (Parameter)	SL (β)	IPD (α)	SL (β) / IPD (α)	SL (β)	IPD (α)	SL (β)	IPD (α)
$\tau^2 = 0$								
Balanced	β_2 or α_2	0.96	0.96	1.00	−0.01 (0.67)	0.00 (0.69)	0.00	0.00
	β_3 or α_3	0.98	0.97	1.00	−0.02 (0.34)	−0.01 (0.34)	0.00	0.00
	β_4 or α_4	0.96	0.96	1.00	−0.01 (0.18)	0.01 (0.18)	0.00	0.00
Unbalanced	β_2 or α_2	0.96	0.96	1.00	−0.02 (0.61)	−0.01 (0.61)	0.00	0.00
	β_3 or α_3	0.94	0.94	1.00	−0.01 (0.35)	0.00 (0.35)	0.00	0.00
	β_4 or α_4	0.93	0.93	1.00	−0.01 (0.19)	0.00 (0.19)	0.00	0.00
$\tau^2 = 0.5$								
Balanced	β_2 or α_2	0.95	0.96	0.98	0.01 (2.27)	0.05 (2.35)	0.01	0.01
	β_3 or α_3	0.96	0.96	0.98	0.00 (1.22)	0.02 (1.23)	0.02	0.03
	β_4 or α_4	0.96	0.96	0.98	−0.02 (0.59)	0.00 (0.60)	0.02	0.02
Unbalanced	β_2 or α_2	0.96	0.96	0.98	−0.04 (2.15)	−0.01 (2.21)	0.01	0.01
	β_3 or α_3	0.97	0.97	0.98	−0.04 (1.11)	−0.02 (1.13)	0.02	0.02
	β_4 or α_4	0.93	0.93	0.99	−0.01 (0.64)	0.01 (0.66)	0.02	0.02
$\tau^2 = 1$								
Balanced	β_2 or α_2	0.96	0.96	0.98	−0.01 (2.81)	0.01 (2.89)	0.02	0.02
	β_3 or α_3	0.95	0.95	0.98	−0.06 (1.60)	−0.03 (1.66)	0.04	0.04
	β_4 or α_4	0.94	0.95	0.98	−0.04 (0.79)	−0.02 (0.81)	0.03	0.04
Unbalanced	β_2 or α_2	0.94	0.94	0.98	−0.06 (3.17)	−0.04 (3.24)	0.03	0.03
	β_3 or α_3	0.94	0.95	0.98	−0.05 (1.66)	−0.04 (1.71)	0.04	0.05
	β_4 or α_4	0.94	0.95	0.98	−0.03 (0.81)	−0.01 (0.83)	0.04	0.04

Table 2, we have comparable coverage near 95% for β_2 in all model fits and for (β_3, β_4), where $\tau^2 \neq 0$. When $\tau^2 = 0$, we have low coverage of β_3 and β_4 in both the SL-C and SL-NC fits, between 70% and 90%. This may be due to $\tau^2 = 0$ being on the boundary of the InvGamma$(0.001, 0.001)$ prior on τ. The low coverage in this particular simulation setting is not a concern for our particular application, since it is highly unlikely for a collection of clinical studies to contain no between-study variability. All MAD ratios are just below 1 with narrow CIWs, showing that posterior sampling variability is still comparable.

Figure 2(A) shows the median and 95% credible intervals for percent bias in the parameter estimates from these three model fits. Percent bias is again at or very near 0, with variance in the percent bias increasing with increasing τ^2 for all considered parameters. In Figure 2(B) we see that both the SL-C and SL-NC fits are more efficient than the IPD fit for each parameter used in the equivalence calculation (as evidenced by the median relative efficiency being greater than 1 for each comparison), a trend that increases with τ^2.

Finally, Figure 2(C) shows the median and 95% credible intervals for the dose toxo-equivalence curves, generated by each model fit, with IPD in pink, SL-C in blue, and SL-NC in orange. Each of these fall along the true curve (in green), and their credible intervals fully overlap on both sides of the estimated median with their width increasing with τ^2. The inclusion of summarized subject-level information $\mathbf{Z_i^a}$ in the SL-C model does not improve the estimation of the dose toxo-equivalence curve compared to the SL-NC model.

Of note is the reduced computing time for the SL model comparing to the IPD model. For example, with a single core of an Intel Sandy Bridge architecture processor using

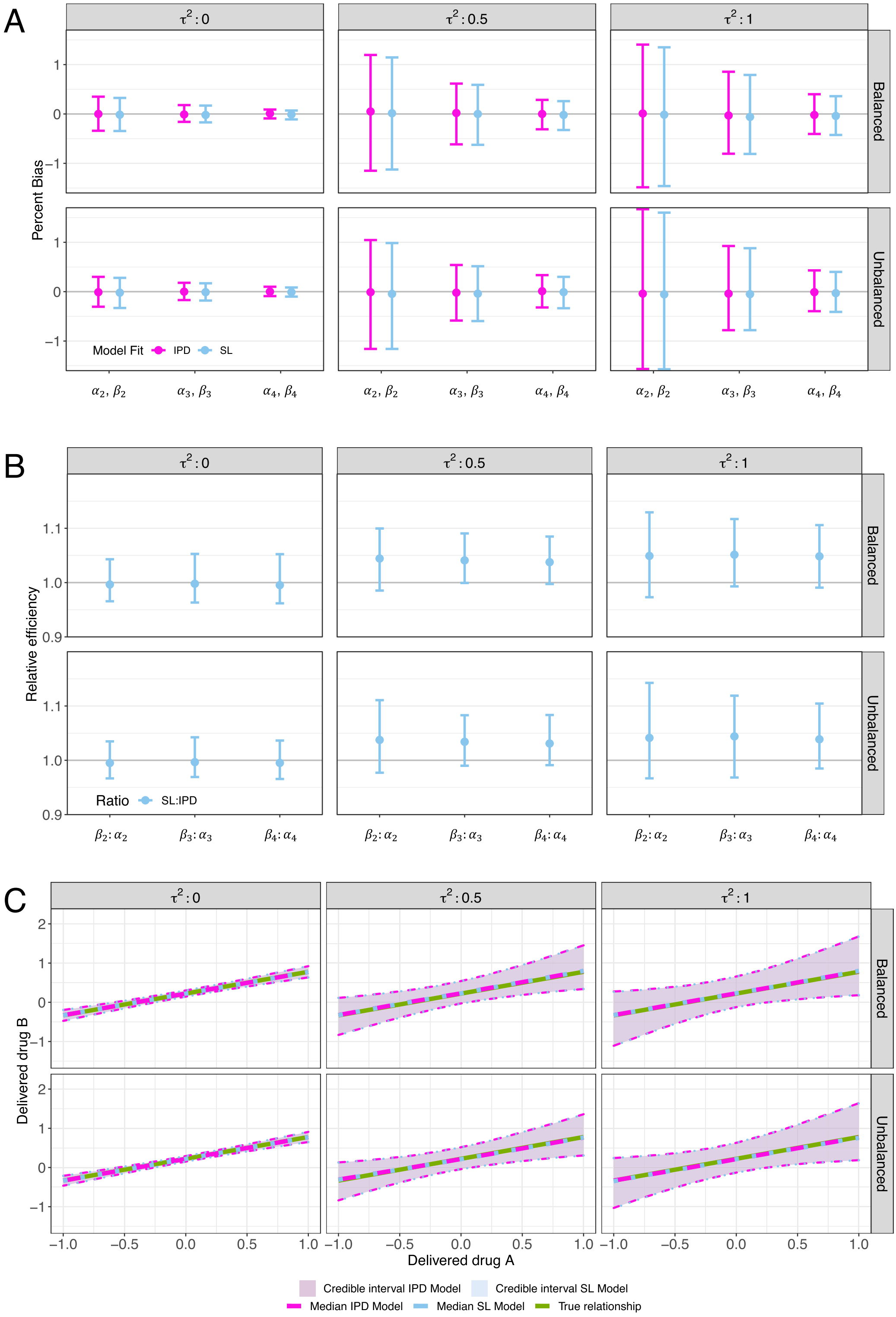

FIG. 1. *Summaries of correctly specified simulations without additional covariates, under each study design and value of* τ^2: (*A*): *Percent bias and* 95% *credible intervals.* (*B*): *Efficiency and* 95% *credible intervals of SL vs. IPD meta models.* (*C*): *Estimated dose toxo-equivalence relationships and* 95% *credible intervals of meta and IPD meta models compared to known relationship.*

TABLE 2

Coverage, median absolute deviation ratios (MADR), percent bias (IQR), and MSE, by fit type and study design, for simulations with additional covariates. Target coverage was 0.95

Simulations With Additional Covariates												
	Statistic	Coverage			MADR		Percent Bias (IQR)			MSE		
Setting	Fit (Parameter)	SL-C (β)	SL-NC (β)	IPD (α)	SL-C (β) / IPD (α)	SL-NC (β) / IPD (α)	SL-C (β)	SL-NC (β)	IPD (α)	SL-C (β)	SL-NC (β)	IPD (α)
$\tau^2 = 0$												
Balanced	β_2 or α_2	0.96	0.95	0.97	0.98	0.98	−0.05 (0.65)	−0.06 (0.62)	0.01 (0.67)	0.00	0.00	0.00
	β_3 or α_3	0.89	0.88	0.95	0.98	0.97	−0.06 (0.35)	−0.06 (0.36)	0.00 (0.37)	0.00	0.00	0.00
	β_4 or α_4	0.78	0.76	0.94	0.98	0.98	−0.06 (0.20)	−0.06 (0.20)	0.00 (0.21)	0.00	0.00	0.00
Unbalanced	β_2 or α_2	0.96	0.95	0.96	0.98	0.98	−0.04 (0.58)	−0.05 (0.59)	0.00 (0.61)	0.00	0.00	0.00
	β_3 or α_3	0.91	0.90	0.98	0.98	0.97	−0.06 (0.29)	−0.06 (0.30)	0.00 (0.31)	0.00	0.00	0.00
	β_4 or α_4	0.73	0.72	0.95	0.98	0.98	−0.06 (0.16)	−0.06 (0.16)	0.00 (0.18)	0.00	0.00	0.00
$\tau^2 = 0.5$												
Balanced	β_2 or α_2	0.96	0.96	0.96	0.94	0.94	−0.12 (2.22)	−0.08 (2.23)	−0.03 (2.38)	0.01	0.01	0.01
	β_3 or α_3	0.96	0.96	0.95	0.94	0.93	−0.06 (1.19)	−0.07 (1.17)	−0.01 (1.25)	0.02	0.02	0.03
	β_4 or α_4	0.89	0.90	0.95	0.94	0.94	−0.08 (0.58)	−0.09 (0.59)	−0.02 (0.63)	0.03	0.03	0.02
Unbalanced	β_2 or α_2	0.95	0.95	0.95	0.94	0.94	−0.07 (2.23)	−0.08 (2.22)	0.00 (2.37)	0.01	0.01	0.01
	β_3 or α_3	0.96	0.95	0.97	0.94	0.94	−0.08 (1.11)	−0.08 (1.12)	−0.02 (1.18)	0.02	0.02	0.03
	β_4 or α_4	0.93	0.92	0.95	0.94	0.94	−0.07 (0.57)	−0.07 (0.58)	−0.01 (0.61)	0.02	0.02	0.02
$\tau^2 = 1$												
Balanced	β_2 or α_2	0.97	0.96	0.97	0.94	0.93	−0.03 (2.79)	−0.01 (2.83)	0.05 (3.05)	0.02	0.02	0.03
	β_3 or α_3	0.93	0.94	0.95	0.94	0.93	−0.06 (1.62)	−0.06 (1.63)	0.00 (1.72)	0.05	0.05	0.06
	β_4 or α_4	0.94	0.94	0.96	0.94	0.93	−0.09 (0.79)	−0.09 (0.80)	−0.02 (0.85)	0.04	0.04	0.04
Unbalanced	β_2 or α_2	0.95	0.94	0.95	0.94	0.94	−0.07 (2.99)	−0.07 (3.11)	0.02 (3.24)	0.02	0.03	0.03
	β_3 or α_3	0.93	0.93	0.94	0.94	0.94	−0.13 (1.63)	−0.13 (1.68)	−0.05 (1.77)	0.05	0.05	0.05
	β_4 or α_4	0.94	0.94	0.95	0.94	0.94	−0.08 (0.75)	−0.09 (0.74)	−0.02 (0.80)	0.03	0.03	0.03

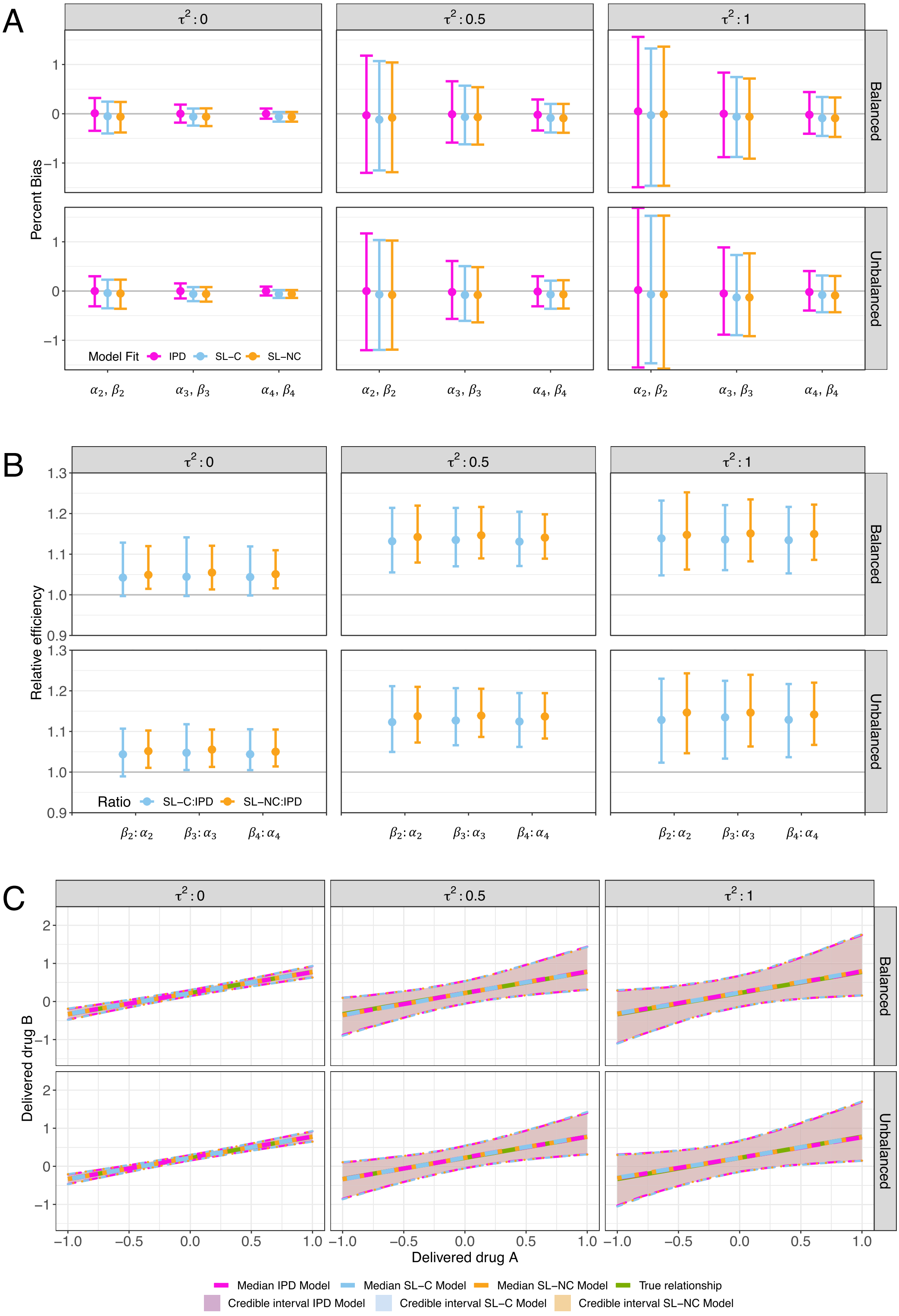

FIG. 2. *Summaries of correctly specified simulations with additional covariates, under each study design and value of* τ^2: (*A*): *Percent bias and 95% credible intervals.* (*B*): *Efficiency and 95% credible intervals of SL-C and SL-NC models vs. IPD meta models.* (*C*): *Estimated dose toxo-equivalence relationships and 95% credible intervals of meta and IPD meta models compared to known relationship.*

two GB of system memory, it took one minute to fit an SL model and five hours to fit an IPD model for a balanced study with no additional covariates and $\tau^2 = 1$. With additional covariates the SL-C and SL-NC models again took one minute, while the IPD model took about 24 hours. Additionally, to check the sensitivity of the method to the simulation coefficients, an alternative set of coefficients chosen to produce similar event rates, $(\alpha_1, \dots, \alpha_6) = (-0.4, 0.1, 0.6, 0.2, 0.3, 0.8)$, were also tested to assess the sensitivity of the model to coefficient choice and produced results consistent with the above simulations in terms of agreement between the IPD and SL/SL-C/SL-NC models.

4.2. *Simulation under a misspecified model.*

4.2.1. *Misspecified model simulation.* We conducted the simulations in Section 4.1 assuming a correctly specified model. Several misspecification scenarios are of interest in evaluating the performance of our proposed method in practice. For example, it is possible that individual patient characteristics, such as age, sex, and related comorbidities (Lutz et al. (2001)), influence their response to medications or general risk for adverse events, possibly in a nonlinear fashion. When this is the case, it is necessary to have complete subject-level data to be able to accurately model these interactions (Debray et al. (2015)). However, of interest is how closely we can approximate the dose toxo-equivalence relationship in the absence of subject-specific information. We explore this first via the inclusion of a nonlinear effect of Z_{ij6} and then via an interaction between Z_{ij5} and either drug type or dose in the data-generating model, comparing the correct specification of the IPD model to the SL-C and SL-NC models fit, as specified in Section 4.1 in the context of aggregated covariates (excluding interactions).

Additionally, our previous simulations assume that all subjects in study i receive the same dose. In practice, however, dosing among subjects can vary due to adverse reactions, missed appointments, or subject noncompliance (Lebovits et al. (1990)). When this is the case, studies will often report an aggregated dose variable, such as the median and IQR of received doses, as in our application in Section 5. To assess the sensitivity of our method to the accuracy of reported doses, we allow for variability in the received dose at the IPD level and fit the SL model on an aggregated dose measure, in this case the median normalized dose. For complete simulation details for these four misspecification scenarios, see the Supplementary Material (Sigworth et al. (2023)).

4.2.2. *Misspecified model results.* Figure 3 shows the estimated dose toxo-equivalence curves for the simulations where subject-level covariates influence the risk of an adverse event in ways that are not considered at the SL-C and SL-NC levels. First, in A we allow the effect of the continuous covariate Z_{ij6} to be nonlinear. Although only the IPD model includes the nonlinear term, we find that both the SL-C and SL-NC models are still able to approximate the true toxo-equivalence relationship, as evidenced by all of the curves overlapping with the true relationship (in green).

Next, in Panels (B) and (C) we consider our simulations where Z_{ij5} is a mediator of the drug or dose response. We treat the binary covariate Z_{ij5} as sex, with $Z_{ij5} = 0$ indicating male and $Z_{ij5} = 1$ indicating female. We display the estimated IPD curves for females (green) and males (yellow), the estimated SL-C (pale blue) and SL-NC curves (orange), and the true curves for females (red) and males (dark blue). In both cases we find the IPD meta model is fairly good at estimating the sex-specific equivalence curves, while the SL-C and SL-NC curves are roughly equal to one another and between the sex-specific curves. Thus, the SL-C and SL-NC curves provide a reasonable estimate of the relationship across sexes but not at a sex-specific level.

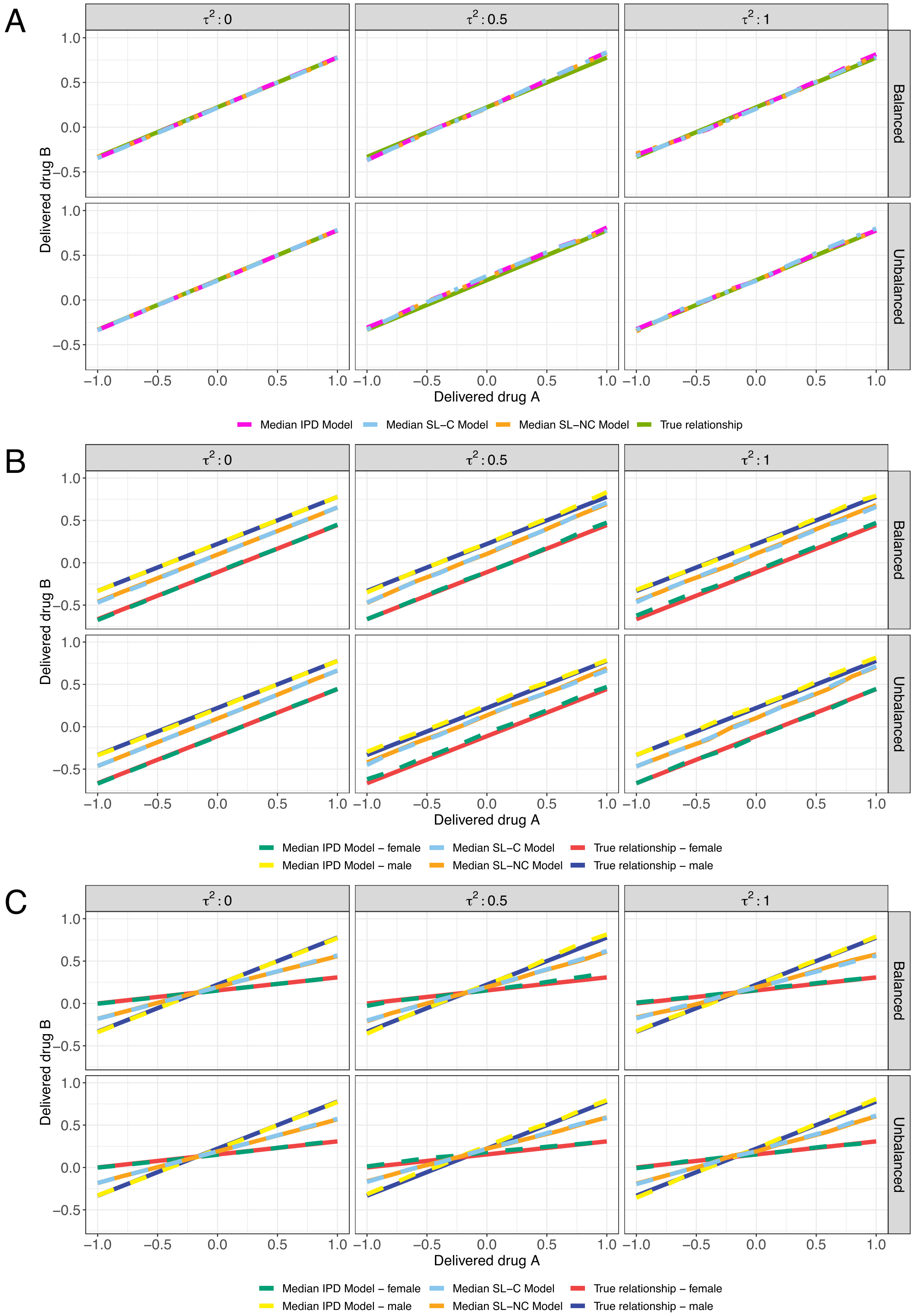

FIG. 3. *Estimated dose toxo-equivalence relationships and 95% credible intervals of SL meta and IPD meta models compared to known relationship with considered excluded term misspecificiations for each study design and value of* τ^2. *(A) Exclusion of nonlinear effect of continuous subject-level covariate. (B) Exclusion of interaction between drug type and a binary covariate (sex). (C) Exclusion of interaction between dose and a binary covariate (sex).*

Finally, the performance of our model when the dose truly varies at the subject level can be found in Figure 4. When allowing dose to vary by subject, we see an increase in variance of the percent bias of β_3 across all simulations (Panel (A)), as compared to the correctly specified models with no covariates in Figure 1(A); however, median percent bias is still at or very near to 0. Additionally, the correctly specified IPD meta model is more efficient than the misspecified meta model for β_3 and β_4 (Panel (B)). Finally, the credible interval for the meta model in the dose toxo-equivalence relationship (Panel (C)) is slightly wider than the IPD meta model, though the median estimated curves are still very similar to both one another and the true relationship.

5. Clinical application. We revisit our clinical application from Section 2 by applying our methods to data from 169 published studies in the taxane chemotherapy drugs paclitaxel and docetaxel. Descriptive summaries of our considered studies can be found in Supplementary Table 1 (Sigworth et al. (2023)). Our outcome Π_i is the observed rate of all-grade neuropathy in each study, and our dose d_i in each study is the normalized median cumulative dose received by subjects, calculated as $d_i = (D_i - \overline{D_i})/(sd_i)$ where D_i is the median cumulative dose received in mg/m^2 in study i, $\overline{D_i}$ is the mean of the reported D_i, and sd_i is the standard deviation of the reported D_i. Dose was normalized within each drug independently (i.e., paclitaxel was normalized with respect to other paclitaxel doses and the same for docetaxel). We consider the inclusion of the study summary variables normalized median age, age_i, and dose gap, dg_i (time between taxane doses in fractions of four-week periods, i.e., $0.25 = 1$ week).

We considered several transformations of D_i prior to normalization, looking to reduce skewness in the distribution of doses. Specifically we looked at $\sqrt{D_i}$, $\ln(D_i)$, and Box–Cox transformations with $\lambda = 0.22$ (chosen to maximize the objective function), $\lambda = 0.25$ (for fourth roots), and $\lambda = 0.33$ (for cube roots, as D_i is a volume). We fit both SL-C and SL-NC models and considered the addition of a random slope to allow the dose-response relationship to vary by study. Each candidate model was fit with a burn-in of 15,000 and 500,000 samples with a thinning interval of 50, across four independent chains, and the deviance information criterion (DIC) was used to compare across models. The DIC values for all models (seven in total) were between 206 and 208 (listed in full in Supplementary Table 2 (Sigworth et al. (2023)), but the lowest and most parsimonious was the fit with covariates, no random slope, and no transformation to dose prior to normalization. Thus, the overall final model structure fit to the data can be defined as follows. Let $\mathbf{X_i} = (X_{iP}, X_{iD}, d_i, \text{age}_i, dg_i)$ be a study-level data vector, and define $Y_i | \mu_i, \boldsymbol{\beta}, \mathbf{X_i} \sim N(\psi_i, S_i^2)$, where $\boldsymbol{\beta}$ denotes the vector of estimated coefficients based on our real data and

$$\psi_i = \mu_i + \beta_1 + \beta_2 X_{iD} + \beta_3 X_{iP} d_i + \beta_4 X_{iD} d_i + \beta_5 \text{age}_i + \beta_6 dg_i$$

with $X_{iP} = I(\text{drug} = \text{paclitaxel})$ and $X_{iD} = I(\text{drug} = \text{docetaxel})$. Noninformative priors of $\mu_i | \tau \sim N(0, \tau^2)$, $\tau \sim \text{InvGamma}(0.001, 0.001)$, and $\boldsymbol{\beta} \sim \text{MVN}(\mathbf{0}, 10^6 \,\text{diag}(\mathbf{1}))$ were set for all model parameters. With the added normalization step for dose D_i, the equivalence relationship for dose D_P of paclitaxel to D_D of docetaxel, in mg/m^2, becomes

$$D_D = \frac{\beta_3(\frac{D_P - \overline{D_P}}{sd_P}) - \beta_2 + \beta_4(\frac{\overline{D_D}}{sd_D})}{\beta_4 / sd_D},$$

where $\overline{D_P}$ is the mean cumulative dose of paclitaxel, $\overline{D_D}$ is the mean cumulative dose of docetaxel, and sd_P and sd_D are the scaling values for paclitaxel and docetaxel, respectively. Once the equivalence relationship was established, the resultant equivalent dose pairs were converted from normalized units to original mg/m^2 units using the centering and scaling values originally used in their normalization.

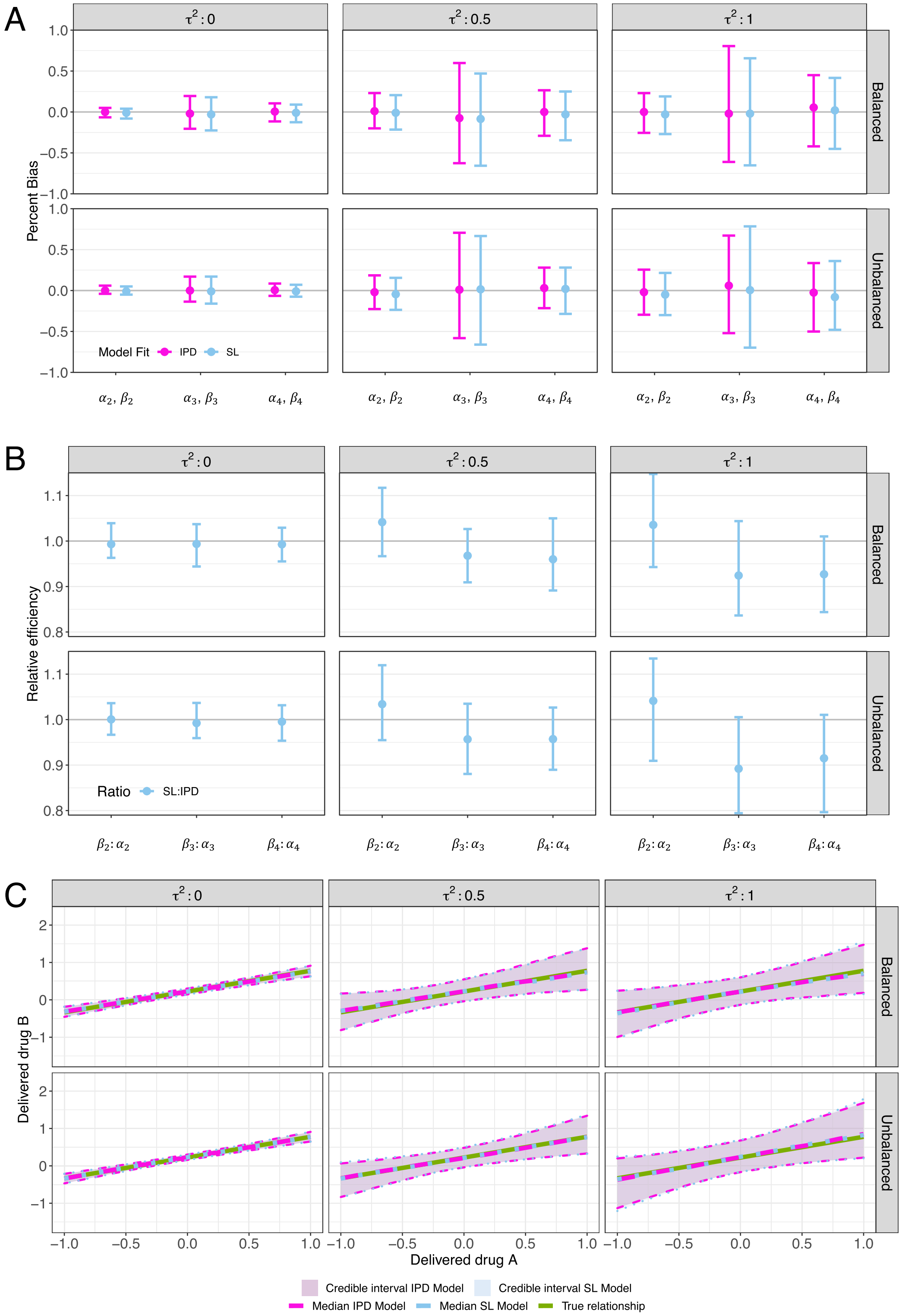

FIG. 4. *Summaries of misspecified simulations with a varied dose, under each study design and value of* τ^2. (*A*): *Percent bias and 95% credible intervals.* (*B*): *Efficiency and 95% credible intervals of SL versus IPD meta models.* (*C*): *Estimated dose toxo-equivalence relationships and 95% credible intervals of meta and IPD meta models compared to known relationship.*

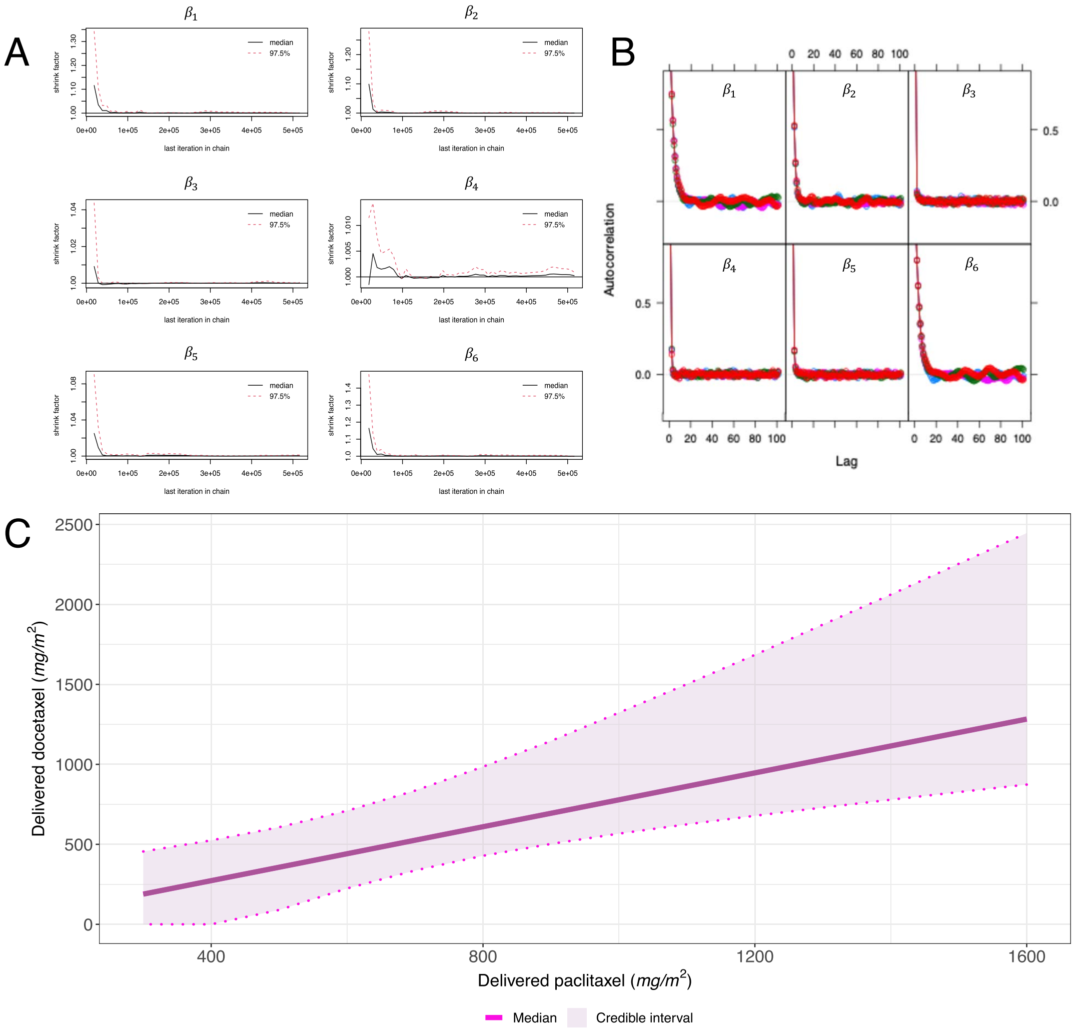

FIG. 5. *Summaries of model fit to taxane data.* (*A*): *Gelman plot.* (*B*): *ACF plot.* (*C*): *Estimated dose toxo-equivalence relationships and* 95% *credible intervals for paclitaxel and docetaxel.*

The final model diagnostics can be found in Figure 5 Panels (A) and (B). The sampling chains converged for each parameter, demonstrated by the Gelman plot in A. There was no evidence of autocorrelation issues in the samples for $(\beta_2, \beta_3, \beta_4, \beta_5)$, demonstrated in the ACF plot for the $\boldsymbol{\beta}$ parameters in B. There is some evidence of autocorrelation up to a lag of 40 with (β_1, β_6); however, as the chain is thinned by 50 and these parameters are not used in the equivalence calculation, this autocorrelation is not an issue.

The dose toxo-equivalence relationship, generated from these samples, can be found in Figure 5(C), evaluated along a range of plausible cumulative paclitaxel doses in mg/m^2. Compared to the simulated curves, the width and shape of the credible interval around this relationship is consistent with our simulation with additional covariates, an unbalanced study design, and higher between-study variability, which is also the most similar simulation design, given the mean estimate for $1/\tau^2$ (an estimate of precision returned by JAGS) was 1.12, resulting in a τ^2 of 0.89. Note the lower credible interval is clipped at 0 mg/m^2, since dosage values must be nonnegative. Along the IQR of paclitaxel doses observed in our data, 656–1085 mg/m^2, the width of the credible interval was roughly equal to the cumulative dose of two treatment cycles of docetaxel, providing useful guidance to clinicians. The lower bound of the credible interval is also particularly informative in a clinical sense, since clinicians can view the lower bound as a cautious dosing threshold from a toxicity perspective. Although we had also hoped to control for cancer type due to different proportions of each cancer represented for each drug, due to sample size constraints, we were unable to do so. As more trial data becomes available, we hope to repeat this analysis while controlling for cancer type.

6. Discussion.

This proposed approach to building a dose toxo-equivalence model from a Bayesian SL meta-analysis consistently produced a good approximation of the true dose toxo-equivalence relationship, performing very similarly to the IPD meta model fit to the full data under all simulation conditions with significantly reduced computing time (e.g., about one minute for the SL meta model as opposed to hours for the IPD meta model). Of note is that, in the absence of an interaction, the relationship between d_A and d_B depends solely on study-level information. This finding is valuable, since efforts to recover individual patient data with clinical trial results are time-intensive, expensive, or even impossible. In estimating the parameters involved in the dose toxo-equivalence relationship, we see no increase in bias or loss of efficiency across conditions, as compared to IPD meta-analysis for those parameters needed to calculate equivalence, an outcome that has been demonstrated in similar works (Luo et al. (2022), Zeng and Lin (2015)). In the case of Zeng and Lin (2015), our approach has two major differences to their work. First, Lin and Zeng (2010a), Lin and Zeng (2010b), Zeng and Lin (2015) were interested in using meta-analysis to combine the parameter estimates from each study/site, with each study including both groups and the parameter being the comparison of two groups, such as an odds ratio, while we are interested in incidence rates in each treatment group with specific doses, and each study can include either one or both groups. Second, adjusting for covariates has a different meaning in their application. In Lin and Zeng (2010a), Lin and Zeng (2010b), Zeng and Lin (2015), the subject-level covariates were adjusted for at the study level before proceeding to meta-analysis, while in our application only the aggregated study-level covariates are available but not the subject-level covariates. In light of these differences, the theoretical conclusions made by Lin and Zeng are not applicable to our empirical findings.

We found no significant improvements in the dose toxo-equivalence estimates when including additional available aggregated covariates in our SL-C model fits, suggesting that potential heterogeneity that may be explained by these aggregated measures does not provide additional information in estimating equivalence, likely because these parameters are not estimating the same quantity between the SL meta and IPD meta models; in extreme cases these parameters could suffer from Simpson's Paradox (Cates (2002), Berlin et al. (2002)). Furthermore, the finding of comparable performance between SL-C and SL-NC was in the absence of an effect modifier. As orthogonality between predictors does not necessarily lead to orthogonality of their coefficient estimates under a logistic regression model (McCullagh and Nelder (1989)), this finding is an empirical observation based on our simulation results. In practice, we could consider both SL-C and SL-NC approaches followed by model selection, however, with no intention to draw connections between the effects of aggregated covariates $\mathbf{Z_i^a}$ and the effects of subject-level covariates $\mathbf{Z_{ij}}$.

Our explorations in Section 4.2 demonstrate that our method is robust to several common types of model misspecification. When a continuous covariate has a nonlinear effect that is overlooked in the SL approach, our method still performs comparably to the IPD model with the appropriate specification. In the case where the drug or dose response is moderated by a binary covariate at the subject level, our method produced an equivalence curve that fell between the curves produced by the two levels of the binary covariate, providing a reasonable approximation of the average relationship in the face of incomplete information. Given that drug and dose responses frequently differ by sex or the presence of comorbid conditions

(Lutz et al. (2001)), this suggests that our method can still provide a useful guide for the dose toxo-equivalence in this context, though there is some loss of information in this context when using an SL-C or SL-NC model, as opposed to an IPD model. Further work would be needed to assess the performance of our method in the case of an interaction with a continuous covariate, such as age or a lab value. Additionally, our simulations did not consider the inclusion of an interaction term in the SL model (with the aggregated $\mathbf{Z_i^a}$); the performance of this model specification is of interest in future investigations. We also found that our model was robust to some variation in the dose received at the subject level, explored at a variation of 10%. This finding supports our model application to paclitaxel and docetaxel, which used the median cumulative dose across each study. However, additional simulations with increasing dose variability could provide more insight into the impact of using an aggregated dose value. Finally, all of the models considered in the simulations and case study assume a linear relationship between transformed dose and rates of neuropathy. An extension on our work would be to explore model performance where the true relationship is nonlinear, which would require constraints on the dose-equivalence calculation for the solution to be identifiable.

Of note is that our simulations and case study focus on nonrare outcomes. The taxane clinical trial data we used had an overall median adverse outcome rate of 0.29 (IQR 0.16–0.48), and the parameters used in our simulations generated similar outcome rates for comparability. Some refinement of our proposed method may be needed to be applicable to studies with rare outcomes, such as the use of integrated likelihood inference (Severini (1998), Berger, Liseo and Wolpert (1999)), a generalized linear mixed model, as opposed to a linear mixed model at the study level to avoid the normal approximation, or the incorporation of Poisson random effects models (Cai, Parast and Ryan (2010)). Additionally, a different prior for the random intercept term μ_i may be helpful in the case of rare outcomes, such as uniform or half-t, as the inverse-gamma prior may incorrectly weight variances when data is sparse (Gelman (2006)). The general concept of using a meta-analytic approach to build a dose toxo-equivalence model in other scenarios remains to be fully investigated, but we believe the findings of our analysis serve as a foundation for other researchers to build upon.

Acknowledgments. Thanks to Leena Choi for discussion on the different misspecification scenarios that we explored. This work was conducted in part using the resources of the ACCRE cluster at Vanderbilt University, Nashville, TN. The authors also wish to thank the Associate Editor and referees, whose comments substantially improved this paper.

Funding. The first author was supported by NIH grant U24CA194215.
The second author was supported by NIH grant T32HG008341.
The third author was supported by NIH grant U24CA194215.
The fourth author was supported in part by NIH grants R01LM012607 and R01AI130460.
The fifth author was supported by NIH grants U24CA194215 and R01CA237895.

SUPPLEMENTARY MATERIAL

Supplementary methods: Misspecification simulation details (DOI: 10.1214/23-AOAS1748SUPPA; .pdf). Supplementary Table 1: Summaries of study characteristics by taxane type. This table summarizes the characteristics of all included studies, stratified by taxane type. Categorical variables are presented as count (%) and compared using a Chi-square test, continuous variables are mean (SD) and compared using a t-test, and discrete numerical variables as median [Q1, Q3] and compared using a Kruskal–Wallis test. Supplementary Table 2: Deviance information criterion values. Deviance information criterion values for all candidate models considered in the application of our method to paclitaxel and docetaxel trial data. This section provides a complete description of our misspecification simulations, including nonlinear effects, drug or dose modifiers, and varying doses.

Supplementary simulation code (DOI: 10.1214/23-AOAS1748SUPPB; .zip). These code files contain all necessary information to reproduce our correctly specified simulations in R.

REFERENCES

ARGYRIOU, A. A., KOLTZENBURG, M., POLYCHRONOPOULOS, P., PAPAPETROPOULOS, S. and KALOFONOS, H. P. (2008). Peripheral nerve damage associated with administration of taxanes in patients with cancer. *Crit. Rev. Oncol. Hematol.* **66** 218–228. https://doi.org/10.1016/j.critrevonc.2008.01.008

BERGER, J. O., LISEO, B. and WOLPERT, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.* **14** 1–28. MR1702200 https://doi.org/10.1214/ss/1009211803

BERLIN, J. A., SANTANNA, J., SCHMID, C. H., SZCZECH, L. A., FELDMAN, H. I. and ANTI-LYMPHOCYTE ANTIBODY INDUCTION THERAPY STUDY GROUP (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. *Stat. Med.* **21** 371–387. https://doi.org/10.1002/sim.1023

CAI, T., PARAST, L. and RYAN, L. (2010). Meta-analysis for rare events. *Stat. Med.* **29** 2078–2089. MR2756556 https://doi.org/10.1002/sim.3964

CATES, C. J. (2002). Simpson's paradox and calculation of number needed to treat from meta-analysis. *BMC Med. Res. Methodol.* **2** 1. https://doi.org/10.1186/1471-2288-2-1

DEBRAY, T. P., MOONS, K. G., VAN VALKENHOEF, G., EFTHIMIOU, O., HUMMEL, N., GROENWOLD, R. H., REITSMA, J. B. and GROUP, G. M. R. (2015). Get real in individual participant data (IPD) meta-analysis: A review of the methodology. *Res. Synth. Methods* **6** 293–309.

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. MR2221284 https://doi.org/10.1214/06-BA117A

HEDGES, L. V. and OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, FL. MR0798597

KRUSCHKE, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, Jags, and Stan*. Academic Press, San Diego.

LAMBERT, P. C., SUTTON, A. J., ABRAMS, K. R. and JONES, D. R. (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *J. Clin. Epidemiol.* **55** 86–94.

LEBOVITS, A. H., STRAIN, J. J., MESSE, M. R., SCHLEIFER, S. J., TANAKA, J. S. and BHARDWAJ, S. (1990). Patient noncompliance with self-administered chemotherapy. *Cancer* **65** 17–22.

LIN, D.-Y. and ZENG, D. (2010a). Meta-analysis of genome-wide association studies: No efficiency gain in using individual participant data. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* **34** 60–66.

LIN, D. Y. and ZENG, D. (2010b). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* **97** 321–332. MR2650741 https://doi.org/10.1093/biomet/asq006

LUO, C., ISLAM, M., SHEILS, N. E., BURESH, J., REPS, J., SCHUEMIE, M. J., RYAN, P. B., EDMONDSON, M., DUAN, R. et al. (2022). DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. *Nat. Commun.* **13** 1–10.

LUTZ, W., LOWRY, J., KOPTA, S. M., EINSTEIN, D. A. and HOWARD, K. I. (2001). Prediction of dose-response relations based on patient characteristics. *J. Clin. Psychol.* **57** 889–900. https://doi.org/10.1002/jclp.1057

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

OLKIN, I. and SAMPSON, A. (1998). Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics* **54** 317–322. MR1626809 https://doi.org/10.2307/2534018

SEVERINI, T. A. (1998). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika* **85** 507–522. MR1665861 https://doi.org/10.1093/biomet/85.3.507

SIGWORTH, E. A., RUBINSTEIN, S. M., CHAUGAI, S., RIVERA, D. R., WALKER, P. D., CHEN, Q. and WARNER, J. L. (2022). Development of a Bayesian toxo-equivalence model between docetaxel and paclitaxel. *iScience* **25** 104045.

SIGWORTH, E. A., RUBINSTEIN, S. M., WARNER, J. L., CHEN, Y. and CHEN, Q. (2023). Supplement to "Building a dose toxo-equivalence model from a Bayesian meta-analysis of published clinical trials." https://doi.org/10.1214/23-AOAS1748SUPPA, https://doi.org/10.1214/23-AOAS1748SUPPB

STEINBERG, K., SMITH, S., STROUP, D., OLKIN, I., LEE, N., WILLIAMSON, G. and THACKER, S. (1997). Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. *Amer. J. Epidemiol.* **145** 917–925.

WARNER, J. L., COWAN, A. J., HALL, A. C. and YANG, P. C. (2015). HemOnc.org: A collaborative online knowledge platform for oncology professionals. *Journal of Oncology Practice* **11** e336–e350.

WHITEHEAD, A. (2002). *Meta-Analysis of Controlled Clinical Trials*. Wiley, New York.

ZENG, D. and LIN, D. Y. (2015). On random-effects meta-analysis. *Biometrika* **102** 281–294. MR3371004 https://doi.org/10.1093/biomet/asv011