Deliverable 4: A simulation study to evaluate the performance characteristics of uncertainty quantification methods of complex models

CONFIRMS Consortium

Niels Hendrickx[1], Xiaomei Chen[1], Gustaf Wellhagen[1], Robin Svensson[2], Andrew Hooker[1], and the CONFIRMS Consortium

[1]Department of Pharmacy, Uppsala University, Uppsala, Sweden
[2] Swedish Medical Products Agency, Uppsala, Sweden

Contents

1. Introduction 3
2. Objectives 3
3. Methods 4
   3.1 Part 1: Evaluation of BHMA 4
      3.1.1 Description of the BHMA model 4
      3.1.2 Core simulation study 6
      3.1.3 Additional simulation scenarios 8
      3.1.4 Frequentist approach 9
   3.2 Part 2: PSA implementation 10
   3.3 Part 3: Use case - exploring how to combine uncertainties 10
      3.3.1 Case Study Design 10
      3.3.2 Workflow 11
   3.4 Software tools 13
4. Results 13
   4.1 Part 1: evaluation of BHMA 13
      4.1.1 Core simulation study – DDI studies 13
      4.1.2 Additional simulation scenarios 20
   4.2 Part 2: Uncertainty propagation 24
   4.3 Part 3: Use case - exploring how to combine uncertainties 24
      Step 1: Simulation of DDI Study Data 24
      Step 2: Parameter Optimization 24
      Step 3: Probabilistic Sensitivity Analysis (PSA) 25
      Step 4: Combining PSA and BHMA Uncertainty 28
5 Discussion 29
   5.1 Part 1: Evaluating BHMA and exploring alternative approach(es) for qualifying PBPK platforms 29
   5.2 Part 2: PSA for analyzing the uncertainty in model input parameters 31
   5.3 Part 3: Exploring approaches for combining uncertainty 31
Disclaimer 32
References 32
Appendix A: Tables of results 35
Appendix B: Codebase used to conduct the simulation study 38
Appendix C: PowerPoint slide deck for dissemination of results 39

1. Introduction

Mechanistic models such as physiologically based pharmacokinetic (PBPK) models have become important tools in drug development, particularly within the framework of model-informed drug development (MIDD) (1). These models enable predictions of drug–drug interactions (DDIs), first-in-human (FIH) dose estimations, etc. In the regulatory aspect, PBPK modeling is commonly used to support decisions related to DDIs, food effects, and dose extrapolation to special populations. One challenge in its application is that PBPK models are characterized by a large number of parameters based on previous studies and literature (2). The uncertainties of those parameters and their impact on the output metrics are difficult to study systematically. A common approach is sensitivity analysis (3), which only investigates how the change of certain parameters affects the relevant output metrics.

Usually, PBPK modeling is performed using modeling platforms, such as Simcyp, GastroPlus, and PK-Sim, which collectively implement mechanistic models based on previous literature and studies for model structure and system-related parameters. In 2018, the European Medicines Agency (EMA) published a Guideline on the reporting of physiologically based pharmacokinetic (PBPK) modelling and simulation (4), which outlines requirements for qualification of PBPK platforms for high-impact regulatory applications, such as waiving clinical DDI studies or informing dosing decisions in specific populations (e.g., pediatric subgroups) based on limited or no clinical data. Certara conducted a Bayesian hierarchical meta-analysis (BHMA) and received EMA's regulatory qualification opinion for Simcyp in 2025 for predicting CYP-mediated DDIs with low to moderate inhibition (5). The BHMA qualification was based on 220 DDI validation datasets (Certara Drug Interaction Database, formerly known as the University of Washington database), evaluating the predictive performance of SimCyp in terms of bias and imprecision of geometric mean ratio (GMR) and between-subject variances (BSV) of AUC_{inf} or C_{max}. However, the methodology and performance of BHMA for mechanistic models in drug development have not been systematically assessed and compared with other meta-analysis approaches. It is unclear how its performance and conclusion are impacted by the design characteristics of the validation datasets, operational features, and model misspecification.

Uncertainty can be understood from two different perspectives. The first concerns the general predictive performance of a complex model platform (such as SimCyp) within a defined range of contexts of use. In this case, uncertainty refers to the imprecision in predicting output metrics of the model platform (e.g., GMR of AUC_{inf}), which partly reflects variations in data quality and model development practices among previous PBPK models. The BHMA method is an example of this approach, as it estimates the prediction bias and imprecision based on the comparison between observed and predicted values of previous DDI studies. The second perspective focuses on the uncertainty in the parameters of a model, such as a drug-specific model developed using a modeling platform (e.g., a PBPK model parameterized for a specific drug based on available research results). In this context, the main interest lies in characterizing the uncertainty of the model parameters and propagating it through the model to understand how it affects the model predictions as well as any secondary, derived parameters of interest. To study the impact of uncertainty in input parameters, the current most common approach is sensitivity analysis, which has several drawbacks, including a lack of probabilistic interpretation, a neglect of correlation between parameters, and challenges in directly informing the predictive performance. Probabilistic sensitivity analysis (PSA) is a method used to assess how the impact of uncertainty in model input parameters affects the output metrics by assigning probability distributions based on prior knowledge (6). Unlike deterministic sensitivity analysis, PSA offers a probabilistic framework for understanding the influence of input uncertainty and facilitates more informed decision-making.

2. Objectives

The objective of this project is to explore and evaluate uncertainty analysis methods for the regulatory evaluation of complex models. Specifically, the project aims:

To explore and evaluate methods for qualifying PBPK platforms in terms of prediction performance (i.e., uncertainty in the output metrics), with a focus on BHMA:

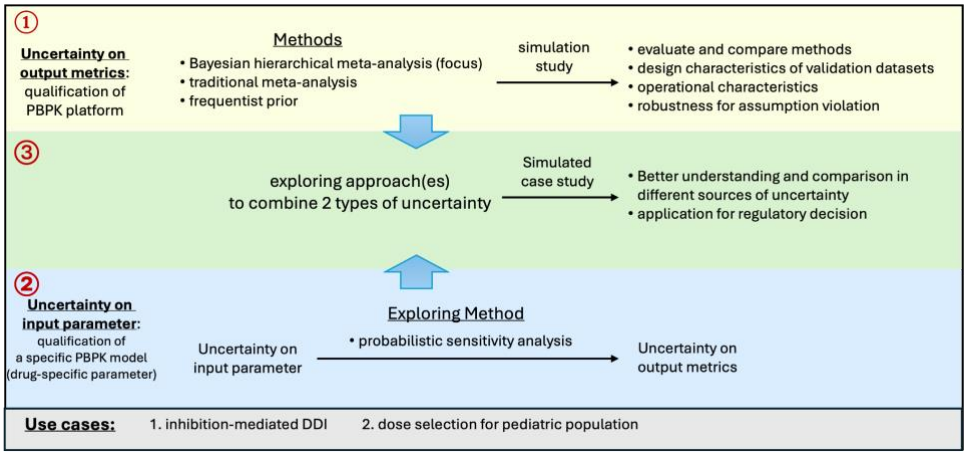
- to investigate the needed design characteristics of validation datasets for BHMA
- to evaluate the operational characteristics of BHMA
- to evaluate the performance of BHMA under a series of scenarios, including model misspecification
- to explore alternative methods, such as traditional meta-analysis

To investigate PSA for characterizing uncertainty in output metrics based on analyzing the uncertainty in model input parameters

- to implement PSA for PBPK modeling with a suggested workflow for its application
- to investigate the impact of selected prior uncertainty distributions on PSA results.

To explore approaches for combining the two sources of uncertainty (uncertainty assessment based on predictive performance and uncertainty in model input parameters)

- to establish a clearer understanding of the different types of uncertainty
- to explore how the investigated approaches can quantitatively inform regulatory decision-making.



3. Methods

3.1 Part 1: Evaluation of BHMA

3.1.1 Description of the BHMA model

The evaluated BHMA model structure is based on the model reported in the Qualification Opinion for the Simcyp Simulator (5). The analyzed dataset comprises 219 CYP-mediated DDI studies. For each study, it includes summary statistics, Geometric Mean Ratios (GMR) and Geometric Between Subject Variance (BSV) of the individual pharmacokinetic concentration-time area under the curve from time 0 to infinity (AUC_{inf}) and maximum concentration (C_{max}) values, reported as both observed (obs) and PBPK-platform-predicted (pred) values, along with sample size and the strength of inhibition.

The BHMA model is shown in Figure 1. For each DDI study i ($i = 1 \dots N_{sim}$), we observe the GMR ($GMR_{i,obs}$) and the BSV $\left(log\ \left(GSD_{i,obs}\right)^{2} = \sigma^2_{sub,i,obs}\right)$.

For the statistical model, the individual GMR is assumed to follow a lognormal distribution. The log-mean is:

$$log\ (GMR_i) = log\ (GMR_{i,pred}) + \beta_{GMR}$$

With $log\ (GMR_{i,pred})$ the log GMR predicted by the PBPK platform for the ith DDI study, and β_{GMR} the bias on the GMR.

The log-variance of that distribution is:

$$\sigma^2_{GMR,i} = \sigma^2_{stu} + \frac{\sigma^2_{sub,i,pred} \times exp\ (\beta_{BSV,i})}{N_{sub,i}}$$

With σ^2_{stu} the between study variance, $\sigma^2_{sub,i,pred}$ the BSV predicted by the PBPK platform for the ith DDI study, $\beta_{BSV,i}$ the bias on the BSV for the ith DDI study, and $N_{sub,i}$ the number of subjects in the ith DDI study.

Then, for the statistical model, the BSV is assumed to follow a gamma distribution, with the shape (k_i) and rate (τ_i) parameters:

$$k_i = \frac{N_{sub,i}}{2}$$

$$\tau_i = \frac{k_i}{\sigma^2_{sub,i,pred} \times exp\ (\beta_{BSV,i})}$$

Finally, a random effect is introduced on the $\beta_{BSV,i}$ parameter, and all parameters have the following priors (either normally distributed or half-Cauchy):

$$\beta_{GMR} \sim N(0,1)$$

$$\beta_{BSV,i} \sim N(\mu_{\beta_{BSV}}, \sigma^2_{\beta_{BSV}})$$

$$\mu_{\beta_{BSV}} \sim N(0,1)$$

$$\sigma^2_{\beta_{BSV}} \sim C(0,1)(R^+)$$

$$\sigma^2_{stu} \sim C(0,1)(R^+)$$

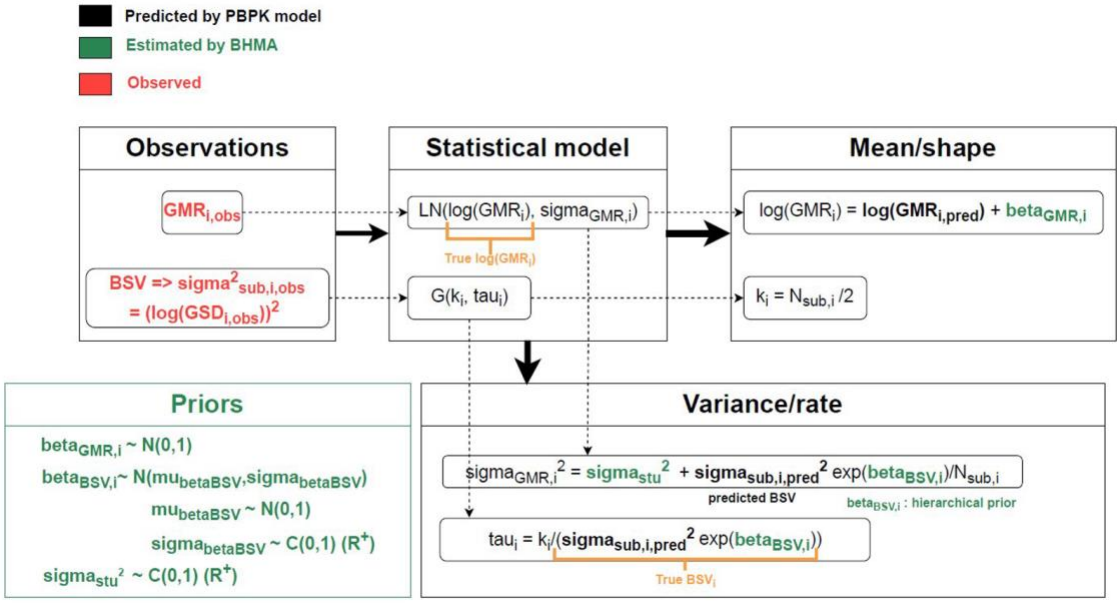


Figure 1: BHMA model and data

3.1.2 Core simulation study

The goal of this simulation study is to assess the performance of the BHMA model in describing the uncertainty of a PBPK platform, and to quantify minimum dataset and study design requirements needed to do so in terms of number of DDI studies, the number of subjects per study, the performance of the PBPK platform, and the estimation settings. This was explored through a series of simulated scenarios. For each scenario (described in 3.2.1), 1000 validation datasets (replicates) are simulated, and the parameters of the BHMA model are estimated. To quantify how uncertainty from the BHMA model propagates into individual PBPK predictions, an additional DDI study, which we call "one case prediction", is simulated for each replicate with a given true GMR. Its predicted credible interval (PCI) is calculated from the BHMA model parameters estimated in that replicate. This allows us to study the influence of the platform uncertainty in the type I error and power for declaring a significant DDI.

3.1.2.1 Simulation scenarios

A simulation scenario is defined as a unique combination of the following parameters: the number of DDI studies per validation set (N_{sim}), the number of subjects per DDI study (N_{sub}), the performance of the PBPK platform (bias and imprecision).

For the simulation study, N_{sim} will have the following values: 4, 10, 20, 50, 200. The number of subjects per DDI study is either equal for all studies within a validation data set (with values of 5, 15, 30) or varied between studies within a validation data set (by sampling from the distribution of number of subjects in the original data (5)). Simulation of the validation data sets is done using the BHMA model described above (see next section for more information). The performance of the PBPK platform is considered using the parameters β_{GMR} (bias), with values of -0.05 (low) or -0.5 (high) and σ_{stu} (imprecision), with values of 0.2 (low) or 1 (high). In total, 80 scenarios are considered.

For each scenario, for each set of validation data simulated, four one case prediction DDI studies are simulated, with a true GMR (see below for a more complete description) of 2 (type I error) or 0.8, 1,

1.5 (power), with the same number of subjects as in the validation data set, and with predicted credible intervals of the following widths: 70%, 80% and 90%.

3.1.2.2 Data

To simulate validation data sets, we need to simulate DDI studies, with predicted and observed GMRs and BSVs. To do so, we need to select a "true" GMR and BSV. In this study we sample observed GMR and BSV in the original data, stratified by inhibitor strength, and set them as the "true" value in the simulated replicate. Since the simulated replicate does not have the same number of DDI studies as in the original data, we need to decide how many studies will be sampled in each inhibitor strength category.

First, we draw the number of DDI studies in each inhibitor strength category by sampling from a multinomial distribution, by using the proportions of each inhibitor strength as observed in the original data (5) as probabilities. Then, for each category of the inhibitor strength, we sample uniformly (with replacement) the corresponding number of studies from the original data. From this sampling, we set the "true" GMR and BSV from the sampled observed GMR and BSV and sample the number of subjects (for the corresponding scenario). Then, we can simulate observed and predicted GMR and BSV from the BHMA model:

$$GMR_{pred,i} \sim LN\left(GMR_{true,i} - \beta_{GMR}, \sigma^2_{stu}\right)$$

$$GMR_{obs,i} \sim LN\left(GMR_{true,i}, \frac{\sigma^2_{sub,i,true}}{N_{sub,i}}\right)$$

$$\sigma^2_{sub,i,pred} \sim LN\left(log\left(\sigma^2_{sub,i,true}\right) - \mu_{\beta_{BSV}}, \sigma^2_{\beta_{BSV}}\right)$$

$$\sigma^2_{sub,i,obs} \sim G\left(k_i = \frac{N_{sub,i}}{2}, \frac{k_i}{\sigma^2_{sub,i,true}}\right)$$

Finally, as mentioned previously, for each replicate, four extra DDI studies are simulated, with a true GMR of 0.8, 1, 1.5 and 2 for the type I error and power assessment.

3.1.2.3 Analysis

For each replicate, the BHMA model is estimated on the simulated validation data set. For the one case predictions, in each replicate, the parameters of the corresponding BHMA model are used to compute predicted credible intervals, by sampling from their posterior distribution to account for parameter uncertainty. For each simulated one case GMR ($GMR_{one,pred}$), the predicted distribution ($GMR_{one,PCI}$) is constructed based on 10000 draws from the posterior distribution, and the statistical model of the predicted GMR (using the bias on the GMR and the between study variance):

$$GMR_{one,PCI} \sim LN\left(log\ \left(GMR_{one,pred}\right) - \beta_{GMR}, \sigma^2_{stu}\right)$$

The predicted credible intervals are calculated as quantiles of the distribution of $GMR_{one,PCI}$. Finally, we declare no significant DDI if the upper bound of the predicted credible interval is below the safety limit (=2), and a significant DDI otherwise.

3.1.2.4 Evaluation metrics

To evaluate the performance of the BHMA model, several evaluation metrics are investigated. First, for each estimated model parameter, the relative bias and relative root mean squared error

(RRMSE) are calculated. Then, to assess the convergence of each parameter, we calculate the fraction of replicates with a $\hat{R}$ below 1.01 (7), the fraction of replicates with an Effective Sample Size (ESS) bulk and tail above 400 (2). Then, to assess the ability of the model to well describe the data, posterior predictive p-values (ppp-value) (8) are calculated for the observed GMR and BSV. The goal is, given a vector of observations y, to simulate observations y^{rep} from the statistical model and calculate $P(y)$, given a summary statistic s (applied across the DDI studies). If the model is well specified, that probability should be 0.5.

To calculate the ppp-value, two methods are proposed in this work, to account for the specific distributions of the statistical model and the hierarchical component of the model. For the first approach, which we call summary statistics ppp-value, the summary statistic is chosen such that, once it is applied to simulated data, the result is independent of model parameters. For the GMR, the Chi square discrepancy measure on the log scale was chosen:

$$s(GMR) = \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} \frac{log\ (GMR_i)\ - log\ \left(GMR_{i,pred}\right)\ -\ \beta_{GMR}}{\sigma_{GMR,i}}$$

If the model is well specified, s(GMR) would follow a χ^2 distribution.

For the BSV, the summary statistic is chosen to standardize its gamma distribution:

$$s(BSV) = \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} \tau_i \times BSV_i$$

If the model is well specified, s(BSV) would follow a χ^2 distribution.

For the second approach, which we call the individual ppp-value, instead of applying a summary statistic across DDI studies, we compute a ppp-value for each DDI study in the validation data set and report the mean individual ppp-value. Finally, for each scenario, we use the extra DDI studies simulated with predicted credible intervals to report the empirical type I error (true GMR=2) and empirical power (GMR=0.8, 1, 1.5) by calculating the fraction of replicates where the upper predicted credible interval of the one case prediction is below 2. For each estimated proportion of empirical power and type 1 error over the 1000 replicates, its 95% confidence interval was constructed using an exact binomial test (using the Pearson-Klopper method).

3.1.3 Additional simulation scenarios

3.1.3.1 Investigation of model misspecification

In addition of the core simulation study, we perform simulations to assess the robustness of the BHMA model to misspecification of the priors of its parameters or to misspecification of the structure of the model. The simulation is performed with low bias and imprecision, varying number of subjects per DDI study and the following number of DDI studies per validation data set: 20, 50 and 200. Several models for the analysis are considered. First, we consider the model used to simulate data, which we call *base*. Then, in one misspecification scenario, we will consider log-normal priors (log-mean=0, log-variance=1) instead of half-Cauchy priors on the $\sigma^2_{\beta_{BSV}}$ and σ^2_{stu} parameters, to consider more informative priors on the variance component, we call this model *LN*. Then, we also consider the misspecification scenarios by changing the standard deviation of the normal priors on the β_{GMR} and $\mu_{\beta_{BSV}}$ parameters, either to an informative prior (standard deviation = 0.1), which we call *inf*, or to a non-informative prior (standard deviation = 5), which we call *ninf*. Finally, we consider

a misspecification on the structure of the model, by removing the random effect on the BSV, which we call *trad*, for traditional meta-analysis.

The data generating mechanism, analysis and evaluation is performed as in the core simulation study. For the evaluation, we consider two additional metrics, based on the differences in Expected Log Predictive Densities (ELPD) (9) between the misspecified model and the base model in each replicate, as well as the fraction of replicates where that difference is greater than 4 points. The ELPD is a metric to support model evaluation and selection, comparable to the objective function value for standard NLME models. It measures the out-of-sample predictive performance of a model, where higher values indicate better performance and an ELPD>4 supports selecting the alternative model (10).

3.1.3.2 Use case 2: Exposure matching

Here, the goal is to evaluate the performance of the BHMA model to quantify the uncertainty for exposure matching, which is a common approach for dose selection in children, based on the GMR of AUC_{inf} with lower and upper boundaries.

In this case study, we assume that the data generating mechanism is identical to the DDI use case, so we can simulate validation data sets based on the original data set comprised of DDI studies. We will also assume that the parameters of the BHMA model and the procedure for building the predicted credible intervals will be identical between the two use cases. Therefore, for the simulation study, the principal difference is the analysis of one case predictions, where we change the safety boundaries to 0.6 – 1.67, and we change the one-sided test to two one-sided tests (11). Therefore, to achieve exposure matching between the two populations (adults and children), we need the lower boundary of the PCI to be above 0.6 and the upper boundary of the PCI to be below 1.67. Otherwise, exposure matching is not achieved. For the evaluation, since nothing changes in the data generating mechanism and the estimation of the BHMA model, estimation metrics will not be considered, and we only consider the type I error and power of declaring exposure matching in one-case predictions.

3.1.4 Frequentist approach

As a complementary analysis, a frequentist approach was also investigated for the power and type I error assessment of one case predictions for DDI studies. For this part, 1000 simulations were performed for each scenario. The three scenarios were 20, 50 or 200 DDI studies, all with random number of subjects. The data generating mechanism was the same as described above, with the GMR set to 1, 1.5 or 2 to assess power and type I error.

The frequentist (nonlinear mixed effects) model was implemented without priors, essentially representing the case with infinitely weak priors. However, the initial estimates were randomly generated from distributions around the true values to be more similar to the Bayesian approach ($\beta_{GMR} \sim \mathcal{N}(-0.0568, 0.5)$, $\mu_{\beta_{BSV}} \sim \mathcal{N}(1.405, 0.5)$, $\sigma^2_{\beta_{BSV}} \sim LN(0,0.3)$, $\sigma^2_{stu} \sim LN(0,0.3)$). The observed GMR_i and BSV_i were set as observations, while the same parameters as in the BHMA model were estimated: β_{GMR}, $\mu_{\beta_{BSV}}$, $\sigma^2_{\beta_{BSV}}$ and σ^2_{stu}.

In frequentist approaches, the simplest and fastest way to assess parameter uncertainty is often through the covariance step, which will then estimate standard errors of parameter estimates. However, to generate confidence intervals of secondary parameters, one must assume the distributional shape of the uncertainty distribution (typically one assumes it to be multivariate normal). Sampling importance resampling (SIR) is a method that allows one to estimate the distributional shape of the parameter uncertainty, and, since it is sampling-based, does not require

any parameter estimation (12). SIR can therefore be more robust but can take more computational time (13).

3.2 Part 2: PSA implementation

The goal of this exercise is to better understand the sources of variability in PBPK models, and whether parameter uncertainty plays a dominant role or not in the uncertainty of the predicted GMR.

Previously identified potential key parameters of the models were investigated via probabilistic sensitivity analysis (PSA). This means that the parameter values are assumed to follow some distribution due to the parameter uncertainty, and different values are drawn from this distribution to reflect uncertainty in GMR.

According to the principle of maximum entropy, the least restrictive distribution compatible that fulfills all constraints should be assumed (14,15). If a mean and variance is known, then a normal distribution should be used. If we also know that the parameter needs to be in the positive domain, the log-normal distribution or truncated normal are possible distributions.

The selected key parameters in this investigation are K_m and k_{cat}, representing the enzyme kinetics following Michaelis-Menten kinetics, where K_m is related to the concentration at which the enzyme reaction rate is half of the maximum and k_{cat} is related to the maximum reaction rate.

3.3 Part 3: Use case - exploring how to combine uncertainties

Part 3 investigates approaches for comparing and combining uncertainty estimates from two complementary sources: BHMA and PSA. The two methods differ in both their focus and their direction of uncertainty quantification. BHMA characterizes platform-level uncertainty by quantifying discrepancies between predicted output metrics and observed DDI outcomes across a collection of historical studies within a defined context of use — that is, uncertainty is assessed at the output level. PSA, by contrast, starts from uncertainty in input parameters, which is then propagated through the model to characterize its impact on predicted output metrics. Input parameter uncertainty can be sourced in different ways, e.g., from uncertainty reported in the literature or from fitting a PBPK model to a particular dataset. Combining these two uncertainty sources aims to provide a more complete picture of prediction uncertainty and to better inform regulatory decision-making for DDI risk assessment.

3.3.1 Case Study Design

Combining uncertainty was explored through a DDI case study reflecting a common regulatory application of PBPK modelling: using a PBPK model for an object of an interaction calibrated against a DDI study with a strong inhibitor to predict the interaction with a weaker inhibitor, thereby supporting a waiver of the corresponding clinical DDI study (4). This scenario falls within the context of use described in the Simcyp qualification opinion (5) covered by the BHMA model applied here.

The following drugs were selected as illustrative examples for the case study:

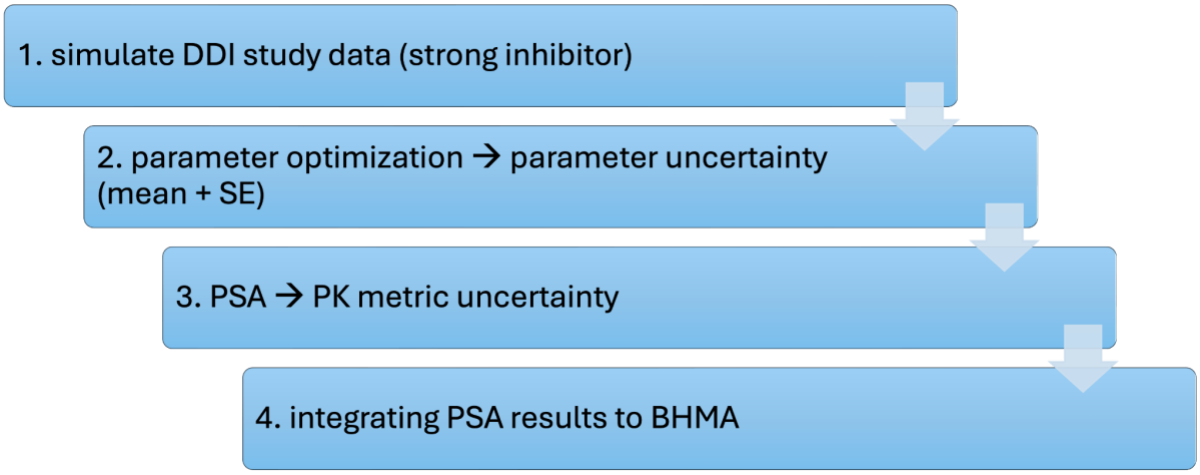
- **Midazolam** — sensitive CYP3A4 substrate (index substrate)
- **Ketoconazole** — strong CYP3A4 inhibitor (used for model calibration)
- **Cimetidine** — weak CYP3A4 inhibitor (prediction target)

All three PBPK models were sourced from the Open Systems Pharmacology (OSP) model library (16) (https://github.com/Open-Systems-Pharmacology/OSP-PBPK-Model-Library), previously published

(17), and implemented in PK-Sim (18). The analyses were conducted in R. Simulations and parameter optimization were performed in PK-Sim with the R package *ospsuite V12.4.2 (19)*.

3.3.2 Workflow

The overall workflow is illustrated below with 4 steps.



Step 1: Simulation of DDI Study Data

A virtual DDI study of midazolam co-administered with ketoconazole was simulated using the OSP midazolam–ketoconazole PBPK model. The study design replicated a published clinical DDI study (20) involving 9 subjects (2 male and 7 female), including the reported dosing regimens.

A virtual population of 9 subjects was generated in PK-Sim, with age and body weight sampled from default distributions bounded by the ranges reported in the source study. This introduced inter-individual variability in physiological parameters (e.g., organ volumes) and in CYP3A4 expression levels, which varied around the reference level defined in the PBPK model. Simulated PK profiles were exported and filtered in R to retain only the time points corresponding to the original clinical sampling schedule.

Step 2: Parameter Optimization

The Parameter Identification function in PK-Sim was used to estimate the CYP3A4 kinetic parameters governing midazolam metabolism: the catalytic rate constant (k_{cat}) and/or the Michaelis constant (K_m). These parameters are central to predicting the magnitude of CYP3A4-mediated DDI.

Optimization was performed separately against two forms of the simulated data from Step 1:

- Mean PK data — the population-averaged concentration-time profile
- Individual PK data — all 9 individual profiles simultaneously

The two approaches are both common practices during PBPK model development.

Step 3: Probabilistic Sensitivity Analysis (PSA)

PSA was conducted using the midazolam–cimetidine PBPK model in PK-Sim (21) via *ospsuite* in R, following the design of a published clinical midazolam–cimetidine DDI study (22).

A parameter space of n = 10,000 samples was generated by random sampling from distributions whose parameters were based on the optimization results of Step 2. Two distributional assumptions were evaluated:

- **Normal distribution:** μ = parameter estimate; σ = standard error from optimization
- **Log-normal distribution:** parameters on the log scale derived from the mean and standard error of the optimized parameter

$$meanlog = e^{\mu+\frac{\sigma^2}{2}}$$

$$sdlog = \left(e^{\sigma^2} - 1\right)e^{2\mu+\sigma^2}$$

For bivariate log-normal distribution, the correlation term is calculated as

$$\rho log = log\left(\frac{\rho\sigma_1\sigma_2}{\mu_1\mu_2} + 1\right)$$

For each of the 10,000 parameter sets, a DDI simulation was run for the typical individual, and the AUC_{inf} ratio and C_{max} ratio (midazolam with vs. without cimetidine) were calculated.

Step 4: Combining PSA and BHMA Uncertainty

BHMA results for AUCinf were taken from the published SimCyp qualification opinion (5). Given that the purpose of this work is methodological exploration and contextualizing the two methods rather than platform benchmarking, a de novo BHMA for PK-Sim was not conducted. As a result, PK-Sim and SimCyp were assumed to have comparable predictive performance for the purpose of this case study.

To compare and add uncertainty from both sources, 90% predictive credible intervals (PCI for BHMA, or predictive confidence interval for PSA) for the AUC_{inf} ratio were constructed under four schemes. Let $\widehat{GMR}$ denote the fixed PBPK-predicted GMR, $GMR_{PSA}^{(k)}$ the *k*-th draw (*k* = 1, …, K) from the PSA-derived AUC_{inf} ratio distribution, $\beta_{GMR}^{(m)}$ and $\sigma_{stu}^{(m)}$ the *m*-th posterior draw (*m* = 1, …, M) of the BHMA bias parameter and between-study standard deviation, respectively, and $Z^{(m)} \sim N(0,1)$ drawn independently for each *m*.

(1) BHMA

$$GMR^{(m)} = exp\left(log\,\widehat{GMR} + \beta_{GMR}^{(m)} + Z^{(m)} \cdot \sigma_{stu}^{(m)}\right)$$

The 90% PCI is $[Q_{0.05}, Q_{0.95}]$ of $\{GMR^{(m)}\}$. This incorporates both BHMA parameter uncertainty (variation in β_{GMR} and σ_{stu} across posterior draws) and between-study variance (the $Z^{(m)} \cdot \sigma_{stu}^{(m)}$ term), and is the method recommended for interval construction in the SimCyp qualification opinion(5).

(2) PSA

$$90\%\ PCI = [Q_{0.05}\left(GMR_{PSA}^{(k)}\right), \quad Q_{0.95}\left(GMR_{PSA}^{(k)}\right)]$$

The 90% PCI is derived directly from the empirical distribution of the K = 10,000 PSA-simulated AUC_{inf} ratios generated in Step 3, without involving the BHMA model.

(3) BHMA + PSA

$$GMR^{(m,k)} = exp\left(log\left(GMR_{PSA}^{(k)}\right) + \beta_{GMR}^{(m)} + Z^{(m)} \cdot \sigma_{stu}^{(m)}\right)$$

The 90% PCI is $[Q_{0.05}, Q_{0.95}]$ of $\{GMR^{(m,k)}\}$. This is identical to scheme (1), except that the fixed $\widehat{GMR}$ is replaced by a draw $GMR_{PSA}^{(k)}$ from the PSA distribution, propagating drug-specific parameter uncertainty into the BHMA framework.

(4) BHMA without between-study variance

$$GMR^{(m)} = exp\left(log\,\widehat{GMR} + \beta_{GMR}^{(m)}\right)$$

The 90% PCI is $[Q_{0.05}, Q_{0.95}]$ of $\{GMR^{(m)}\}$. By omitting the $Z^{(m)} \cdot \sigma_{stu}^{(m)}$ term, this retains only the bias parameter and its uncertainty, excluding between-study variance. This scheme was included to investigate the relative contribution of each uncertainty component, i.e., to compare the magnitude of between-study variance against bias parameter uncertainty.

3.4 Software tools

Data exploration and visualization was performed in R version 4.4 or higher (23). The simulation study was performed with the R package *SimDesign* version 2.22 (24). Parameter estimation was performed using the *Stan* software version 2.37 (7), using the R package *cmdstanr* version 0.9.0 (25). The estimation was carried out with 4 chains, 2500 warmup iterations and 2500 sampling iterations. For the estimation with the frequentist approach, NONMEM 7.6 with the Laplacian estimation method was used (26). To assess the parameter uncertainty, either the covariance step with preconditioning (27) was used, or sampling-importance resampling (SIR) (12). The models with SIR were executed with NONMEM 7.5. Simulations with PBPK models were performed using PK-Sim (v12.2.112) (18) and the *ospsuite* R package (16).

4. Results

For the results of the simulation study, the metrics on the $\hat{R}$ and ESS can be found in the supplementary tables 1. For type I error and power assessment, we only show results with a true GMR of 1, 1.5 and 2, with 90% predicted credible intervals, the rest can be found in supplementary tables.

4.1 Part 1: evaluation of BHMA

4.1.1 Core simulation study – DDI studies

As shown in supplementary table 1, the $\hat{R}$, ESS bulk and tail reach their target value in more than 99% of replicates in all scenarios, meaning that the posterior distribution of the mean is reliable. In Figure 2, we see the relative bias of all parameters for each scenario. The bias is high (>10%) or very

high (>30%) for most parameters with 4 and 10 DDI studies per validation data set. Then, the bias is high for some parameters with 20 DDI studies, but it is below 10% for all other scenarios.

In Figure 3, we see the results for the RRMSE for all parameters. We notice that it is very high (>70%) for the β_{GMR} and σ^2_{stu} parameters with less than 20 DDI studies per validation data set. The other two parameters have high (>30%) RRMSE with 4 and 10 DDI studies per validation data set. The RRMSE is low for other scenarios. We also see no noticeable impact of the number of subjects per DDI study and bias magnitude on the RRMSE.

In Figure 4, the fraction of replicates with a Summary Statistic ppp-value between 0.25 and 0.75 for the GMR and BSV is presented. We see that the ppp-value reaches its target in 100% of replicates for all scenarios for the BSV. However, for the GMR, with low between study variance, it starts at 50% with 4 DDI studies per validation data set and increases with the number of DDI studies. The low ppp-values for GMR with low number of studies can be attributed to the influence of high variability in σ^2_{stu}. In Figure 5, we see the fraction of replicates with an Individual ppp-value between 0.25 and 0.75 for the GMR and BSV. We can see that, for the GMR, the ppp-value reaches its target in 100% of replicates. However, for the BSV, the metric increases with the number of DDI studies per validation data set. This may be due to the stochastic element when only a low number of studies are used. We also notice no impact of the magnitude of the between study variance, the bias magnitude, and the number of subjects per DDI study on the ppp metric.

In Figure 6, results are shown for the type I error and power of declaring no significant DDIs for one case predictions, with true GMRs of 2, 1 and 1.5, and 90% PCIs. First, we see that the type I error is controlled for all but one scenario. We also see that the magnitude of the bias has no effect on the power and type I error. It is apparent that the magnitude of the between study variance has no effect on the type I error, but it greatly decreases the power in all scenarios. With 4 DDI studies, the power is very low (<10%) in all scenarios, and the power increases with the number of DDI studies per validation data set, with a plateau at 20 DDI studies per validation data set. Finally, the number of subjects per DDI study has an effect only with 10 DDI studies per validation data set or less.

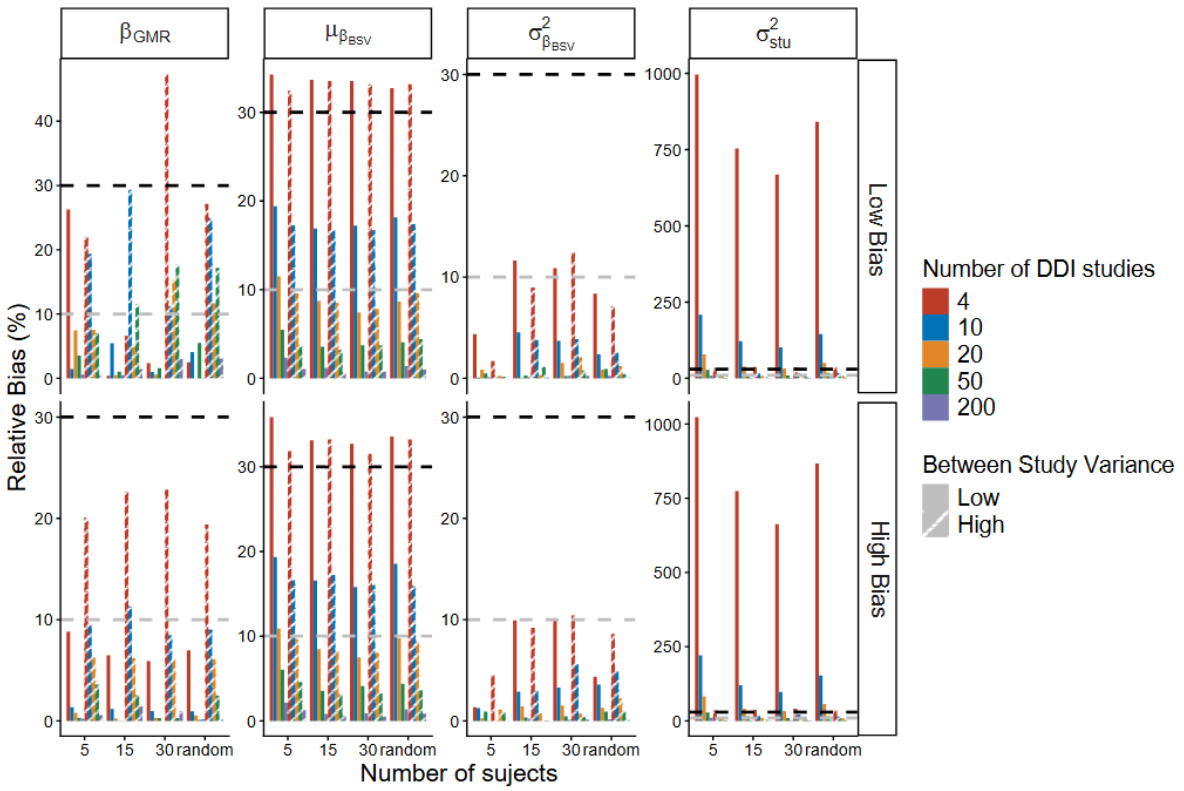


Figure 2: relative bias (%) for all the parameters of the BHMA model in all scenarios for the core simulation (DDI studies). The dashed grey line represents 10% bias, and the dashed black line represents 30% bias

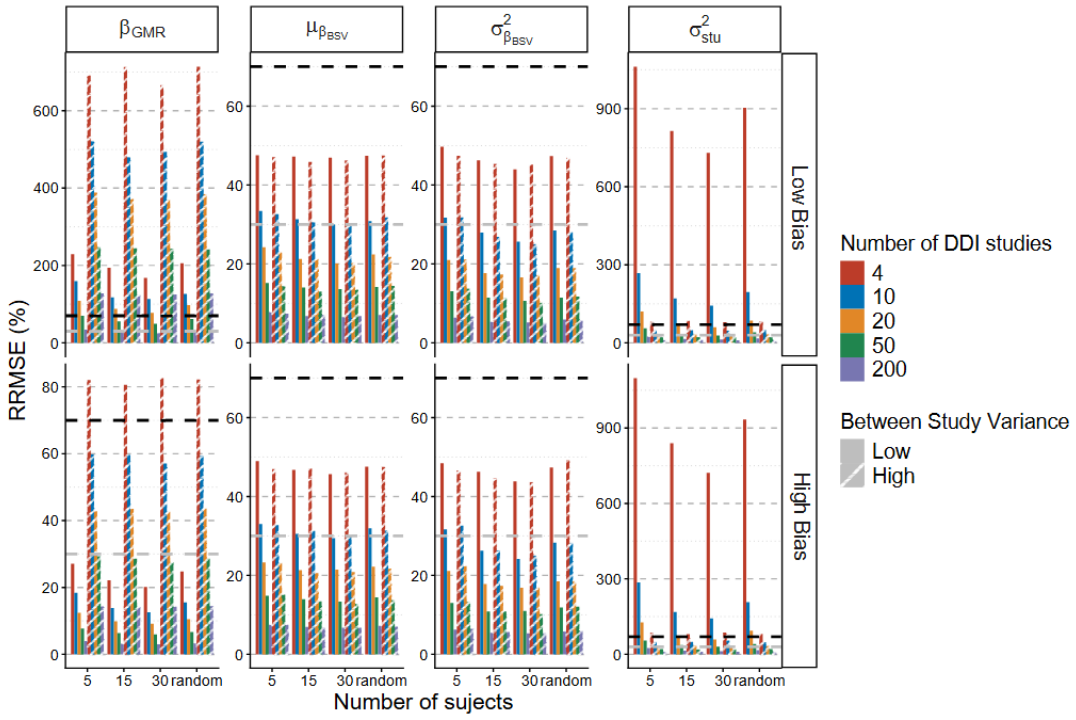


Figure 3: relative RMSE (%) for all the parameters of the BHMA model in all scenarios for the core simulation (DDI studies). The dashed grey line represents 30% RRMSE and the dashed black line represents 70% RRMSE

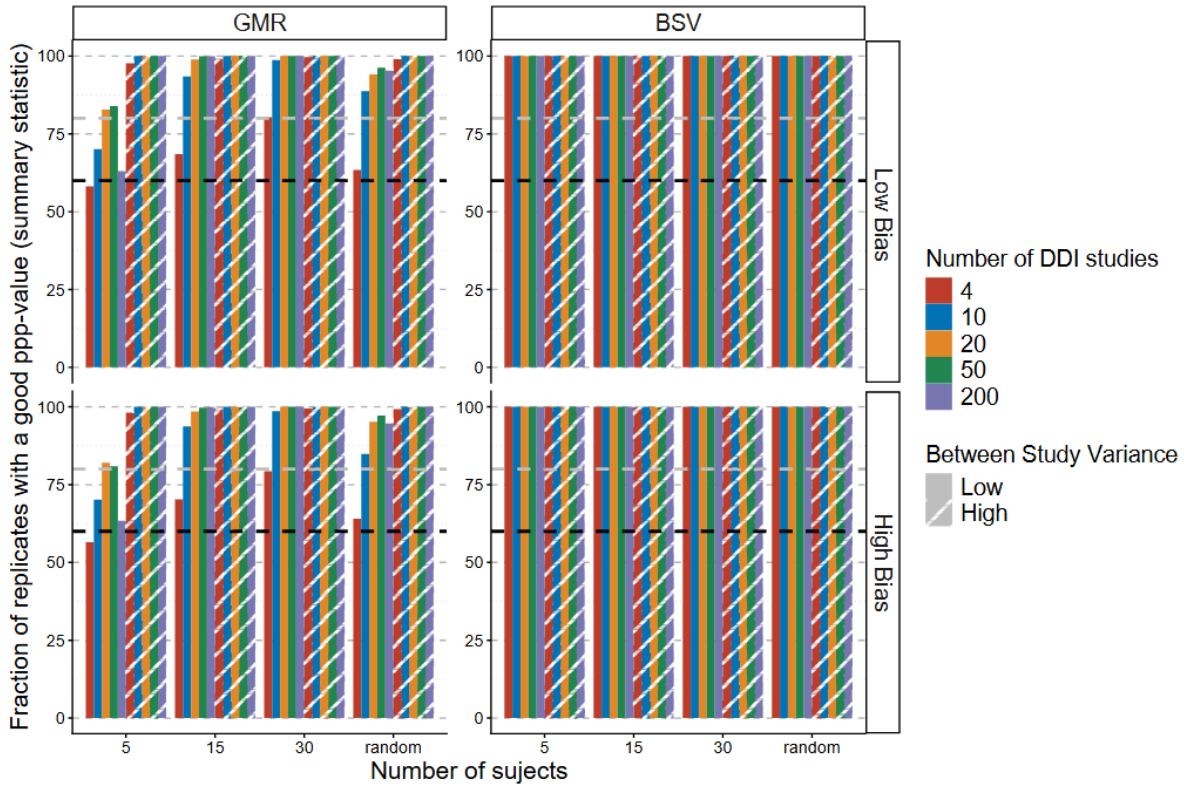


Figure 4: fraction of replicates with a "Summary Statistic" ppp-value between 0.25 and 0.75 for all scenarios in the core simulation.

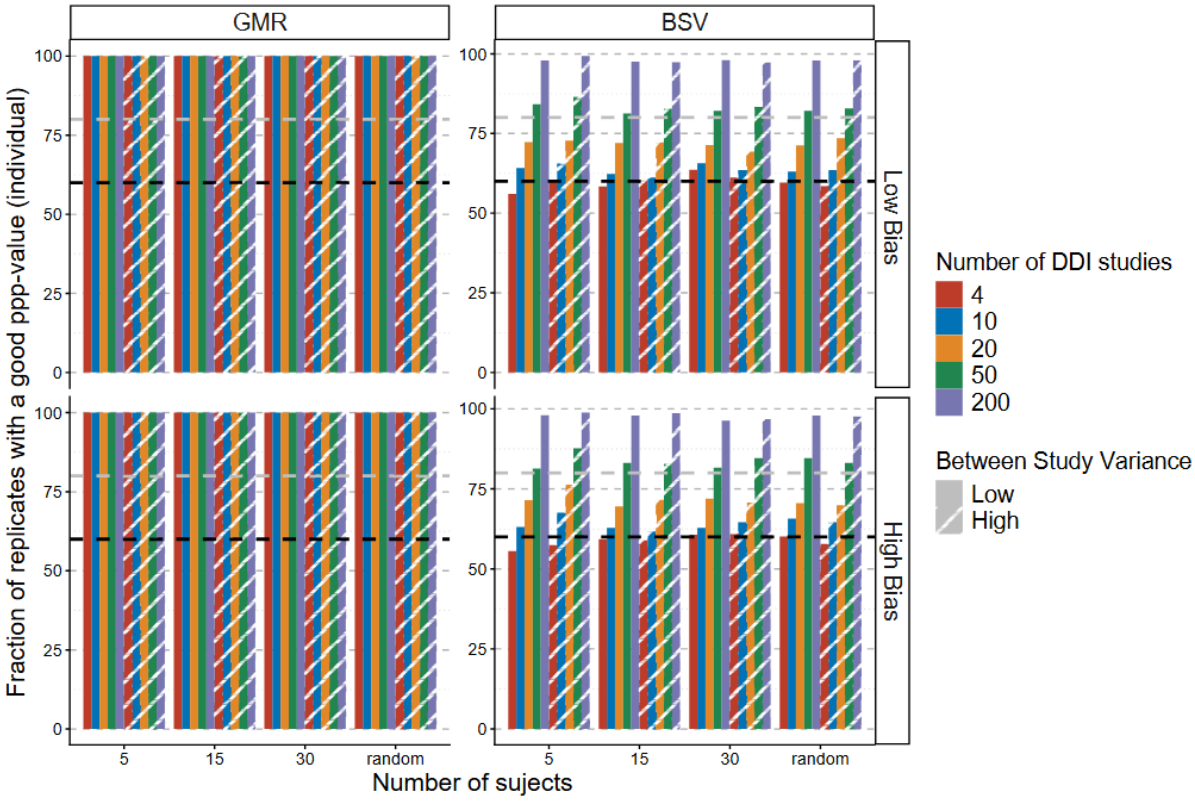


Figure 5: fraction of replicates with an "Individual" ppp-value between 0.25 and 0.75 for all scenarios in the core simulation

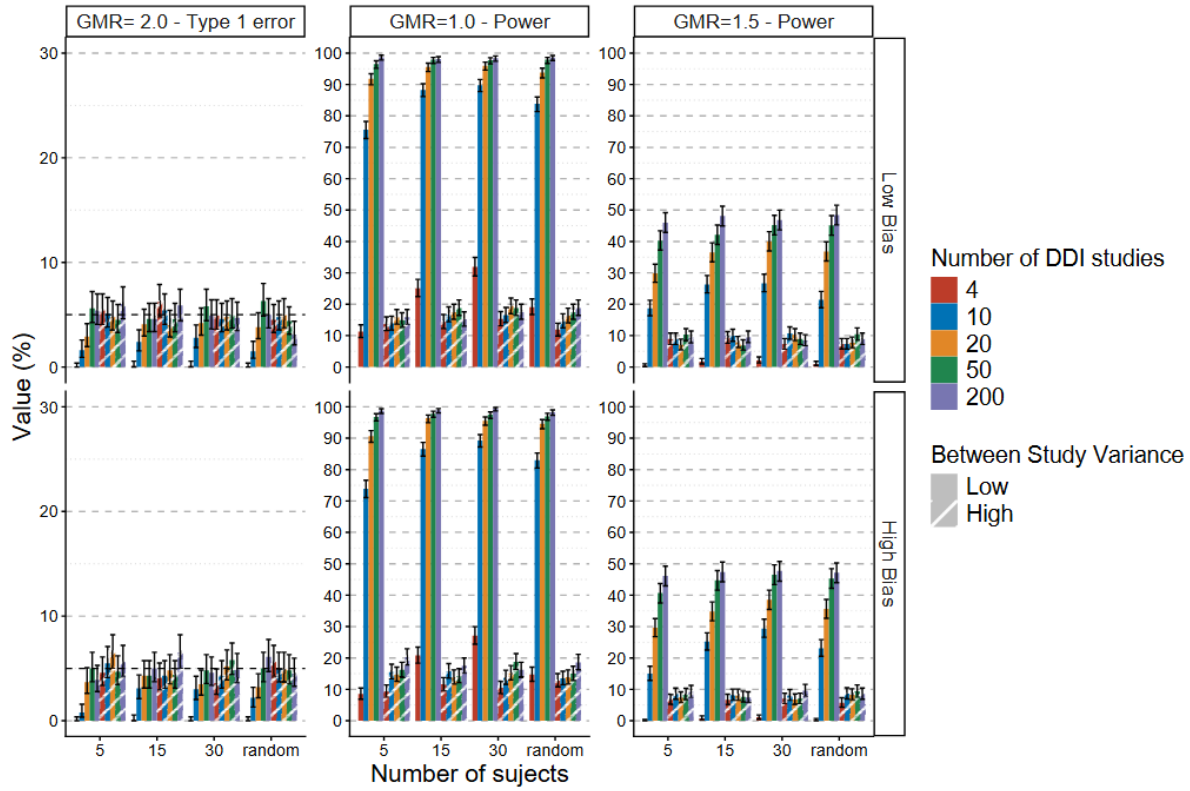


Figure 6: type I error and power to declare no DDIs in one case predictions with using 90% PCIs in all scenarios for the core simulation

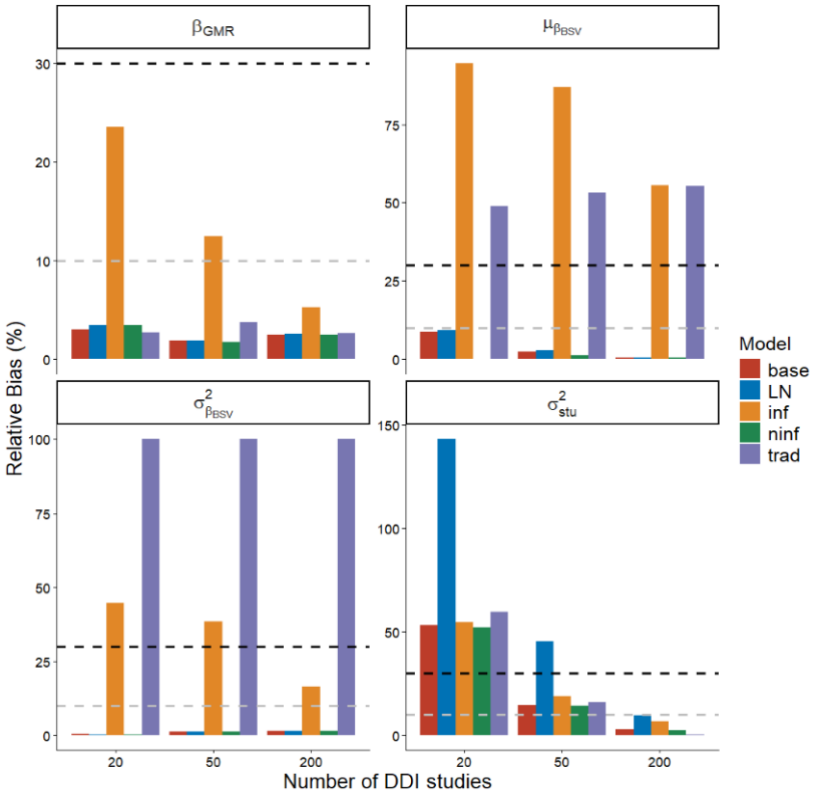


Figure 7: relative bias (%) of all parameters of the BHMA model for all misspecified scenarios. The grey dashed line represents 10% bias, and the black dashed line represents 30% bias.

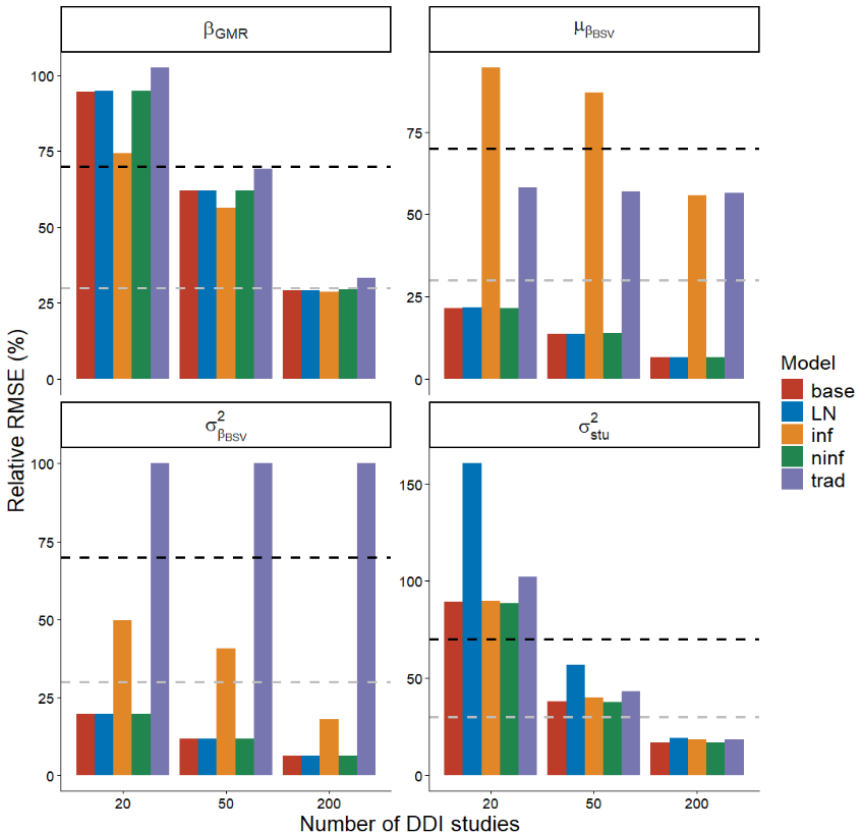


Figure 8: relative RMSE (%) of all parameters of the BHMA model for all misspecified scenarios. The grey dashed line represents 30% RRMSE and the black dashed line represents 70% RRMSE

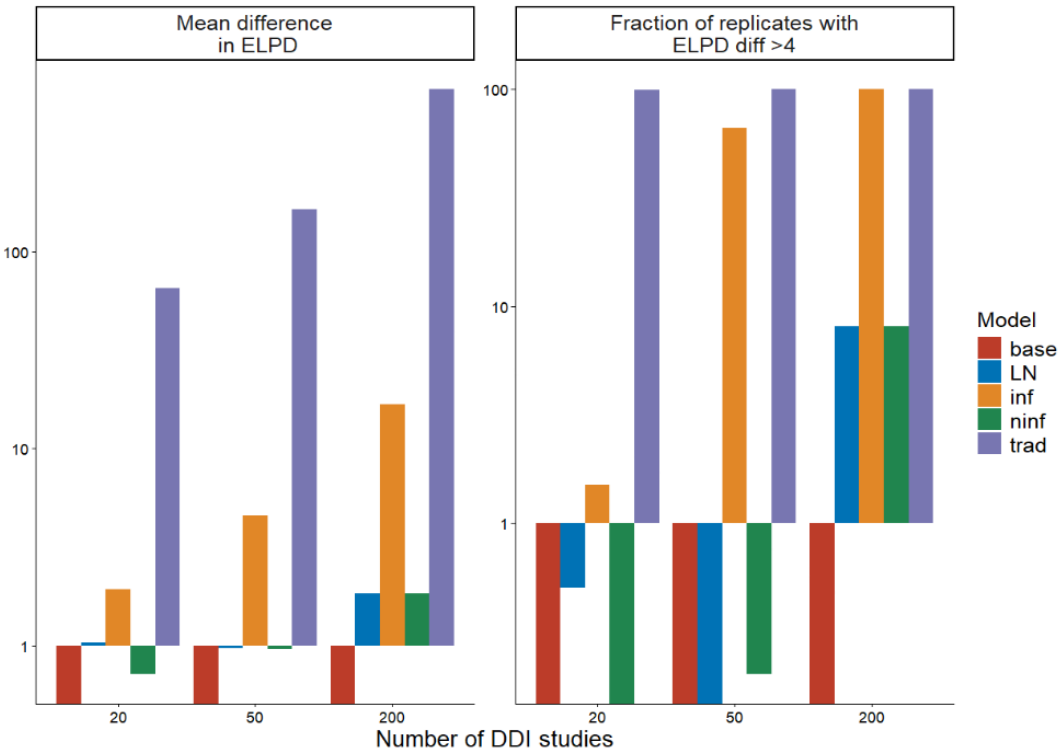


Figure 9: mean difference in ELPD between a misspecified model and the base model (left) and the fraction of replicates where the difference in ELPD is greater than 4 for all misspecified scenarios.

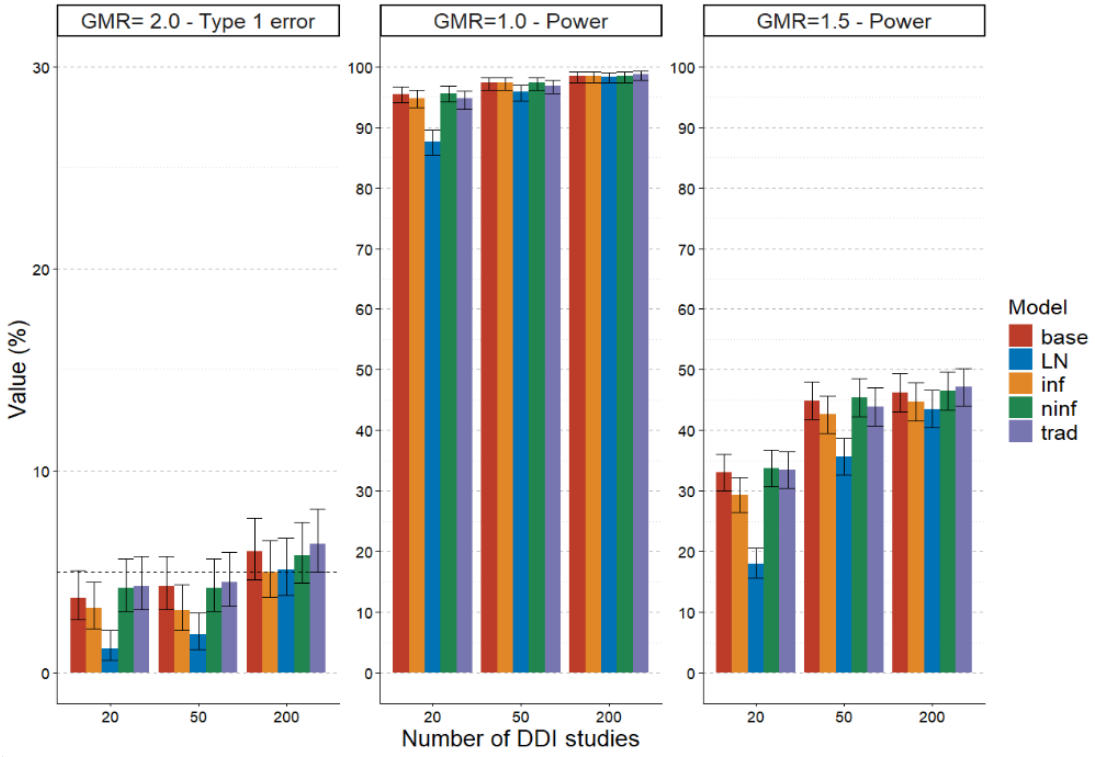


Figure 10: type I error and power to declare no DDIs in one case predictions with 90% PCIs in all misspecified scenarios

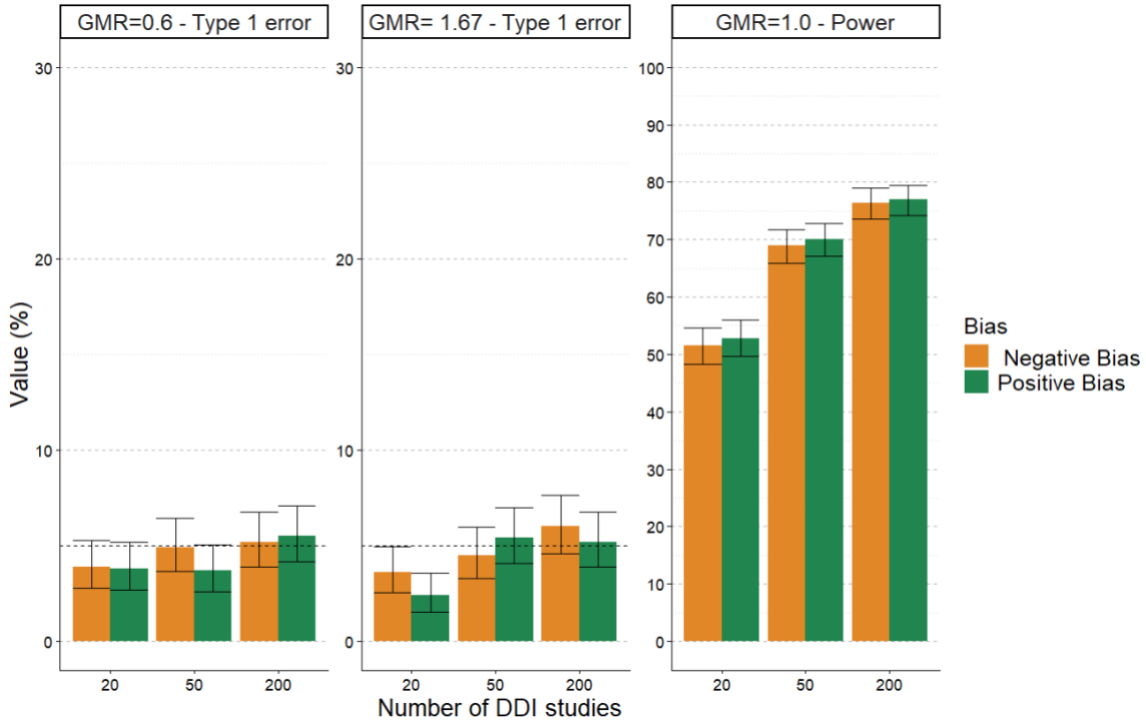


Figure 11: type I error and power to achieve exposure matching in one case predictions with 90% PCIs in all exposure matching scenarios

4.1.2 Additional simulation scenarios

4.1.2.1 Model misspecification

In this work, we now look at the results of the scenarios of model misspecification and exposure matching. In Figure 7, we see the graph of the relative bias of all parameters of the BHMA model for all misspecification scenarios. We notice that the *inf* model (informative priors) has high bias on most parameters, especially with lower number of DDI studies per validation data set. Then, we see that the *ninf* model, with non-informative priors, has the same bias as the base model on all parameters. For the *trad* model, without random effect on the BSV has the same bias as the base model for the β_{GMR} and σ^2_{stu} parameters, and it has high bias for the $\mu_{\beta_{BSV}}$ parameter. For the *LN* model, with lognormal priors on the variance components, the bias on the σ^2_{stu} parameter is very high with 50 DDI studies per validation data set or less, otherwise, the bias is comparable to the base model. In Figure 8, the RRMSE of the parameters of the BHMA model for the misspecification scenarios are presented. The results are in accordance with the results from the relative bias. Figure 9 compares the mean difference in ELPD between a misspecified model and the base model (left), and the fraction of replicates where that difference is greater than 4 (right), which is considered a cutoff for model selection. First, we see that the mean difference in ELPD is very high with the *trad* model and high for the *inf* model, with a difference increasing with the number of DDI studies per validation data set. The mean difference in ELPD is low for the other models. On the right panel, we see that both the *LN* and *ninf* models are not significantly different (ELPD>4) from the base model in most replicates, contrary to the *LN* and *inf* models. In Figure 10, the results for the power and type I error of declaring no DDIs for one case simulations with 90% PCIs in all misspecification scenarios are shown. We first notice that the type I error is controlled with all misspecified models, and very low type I error for the *LN* model with 50 DDI studies per validation set or less. The results show that, except for the *LN* model, all other models have comparable power.

4.1.2.2 Exposure matching scenarios

For the exposure matching scenarios, Figure 11 shows the results for the power and type I error of achieving exposure matching for one case predictions using 90% PCIs. The type I error is controlled on both safety boundaries, at GMR=0.6 and GMR=1.67. In addition, the sign of the bias has close to no impact on the type I error and power. Finally, the power to achieve exposure matching with a true GMR of 1 starts at 50% with 20 studies per validation data set and goes up to 78% with 200 studies. Since it may be rare to have 200 pediatric exposure matching studies, the likelihood for achieving exposure matching under these conditions is low.

As can be seen in Figure 12, the relative bias of the frequentist model was different compared to the BHMA model in several respects: the β_{GMR} parameter had lower bias with 50 validation studies, but higher with 20, however the bias was overall low. The σ^2_{stu} parameter had much lower bias with the frequentist model, and in the opposite direction. In Figure 13, we see the RRMSE, which was generally similar in all cases, and only lower for the frequentist approach on the σ^2_{stu} parameter, especially for 20 validation studies.

The one case predictions investigated with frequentist models showed similar power and type I error properties as the BHMA, see Figure 14. The type I error rate was inflated when looking at only 20 or 50 validation studies, which led us to explore SIR as a method to quantify the parameter uncertainty for that case. With SIR, the type I error rate was controlled for 50, but still elevated for 20 validation studies. The power was similar to that of the BHMA when looking at 50 or 200 validation studies, but higher for the frequentist approach in the case of 20 validation studies.

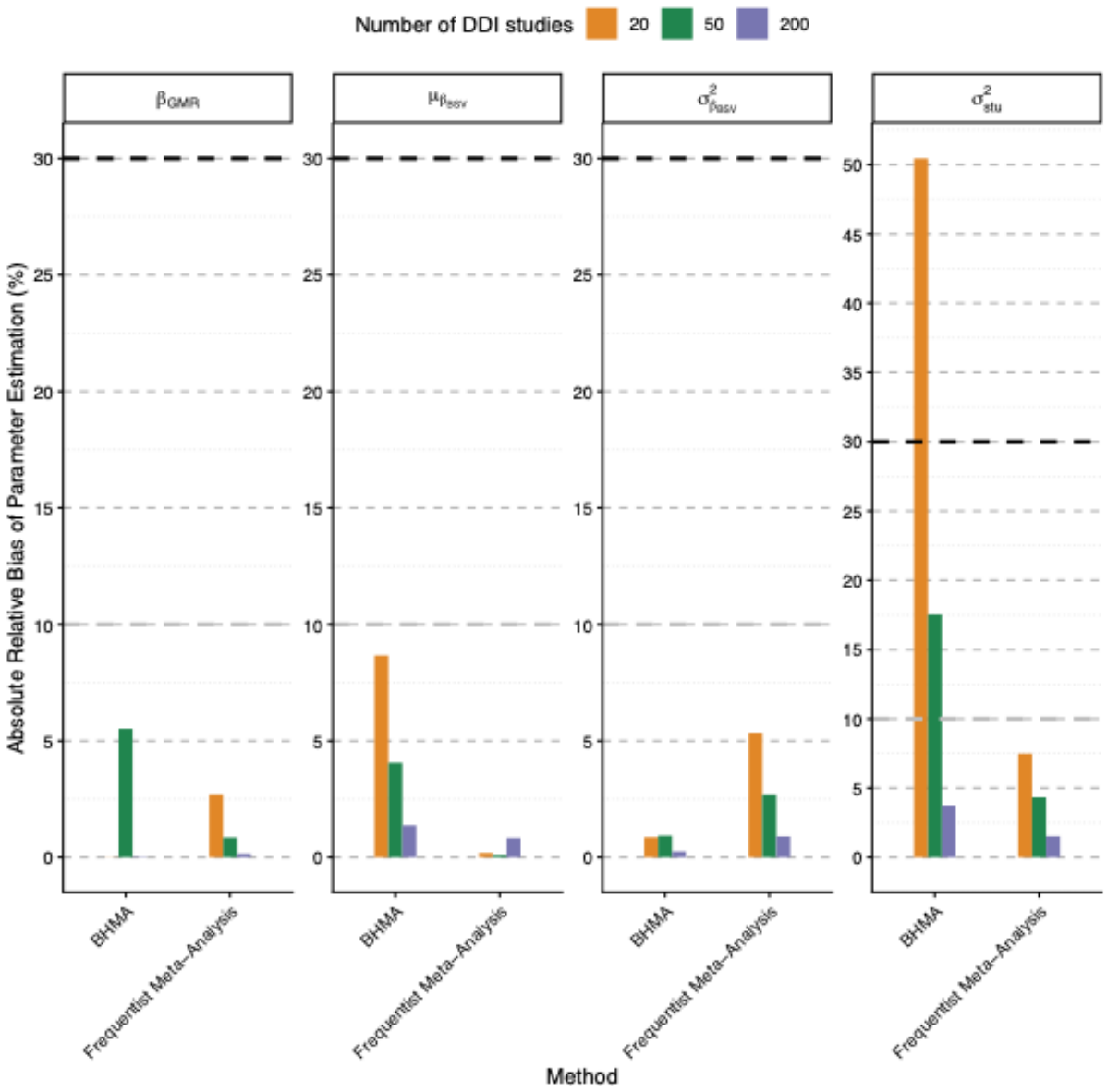


Figure 12: Graph of the absolute relative bias (%) of all parameters of the frequentist and BHMA model for the base scenario. The grey dashed line represents 5% bias.

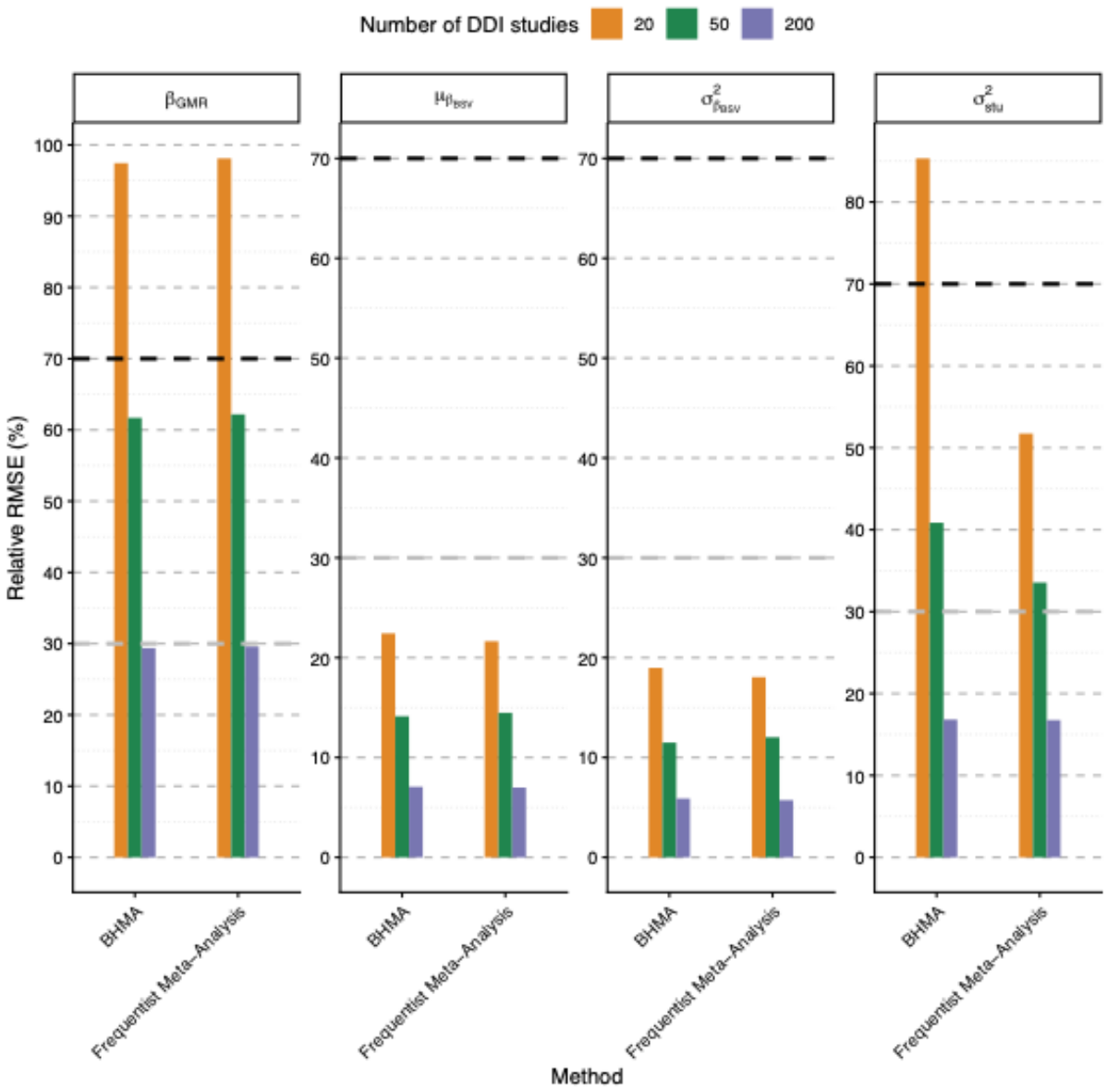


Figure 13: relative RMSE (%) of all parameters of the frequentist and BHMA model for the base scenario. The grey dashed line represents 30% RRMSE and the black dashed line represents 70% RRMSE.

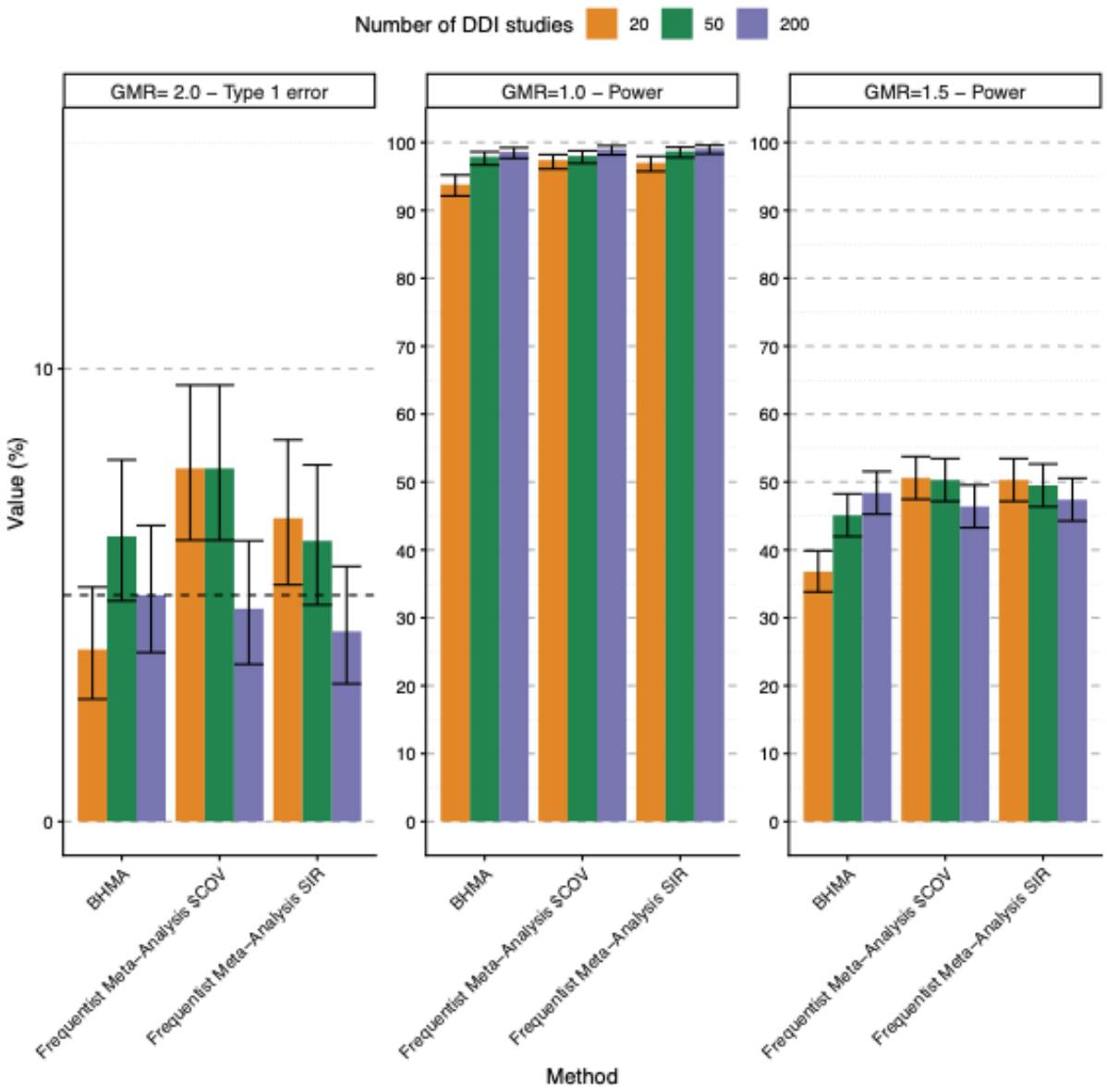


Figure 14: type I error and power to declare no DDIs in one case predictions with 90% PCIs with different estimation methods compared.

4.2 Part 2: Uncertainty propagation

According to the principle of maximum entropy, the most appropriate distribution is the one that puts the least restrictions on the variable, while still fulfilling any constraints, in our case based on previous knowledge about a particular parameter. If only the mean and variance is known, the normal distribution should be used. If only the minimum and maximum value are known, a uniform distribution should be used, and if in addition the mode is known a triangular distribution is appropriate. In our case we have the additional constraint of a positive parameter value to consider, such that log-normal or truncated normal could be used if the variability of the parameter is likely to cause samples below 0.

In this study, we used PK-Sim to simulate the resulting PK metric by varying the value of either k_{cat}, K_m or both. In the next section (Part 3), we show how uncertainty propagation can work if parameters are optimized based on clinical studies, both in a single-parameter and a two-parameter case, and how to combine the input parameter distributions to assess the resulting metrics.

4.3 Part 3: Use case - exploring how to combine uncertainties

Step 1: Simulation of DDI Study Data

Concentration-time profiles for the simulated DDI study (9 virtual subjects) are shown in *Figure 15*. Co-administration of ketoconazole, a strong CYP3A4 inhibitor, markedly increased midazolam exposure, as reflected by elevated AUC_{inf} and C_{max} values relative to midazolam alone. This is consistent with the expected pharmacokinetic interaction and confirms that the simulated data capture the inhibitory effect required for downstream parameter optimization.

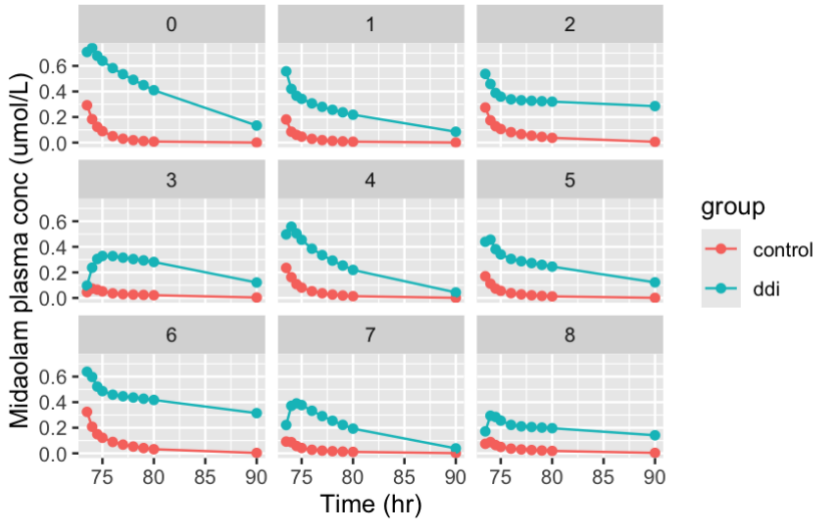


Figure 15. Simulated concentration-time profiles for midazolam with and without ketoconazole in 9 virtual subjects.

Step 2: Parameter Optimization

Parameter optimization results are summarized in *Table 1*. Parameter estimates differed across optimization strategies, both by input data type (mean PK vs. all individual PK data) and by the number of parameters optimized simultaneously (k_{cat} only, K_m only, or both).

When only a single parameter was optimized, the standard error (SE) was small and comparable between the two data types. When both k_{cat} and K_m were optimized simultaneously, SEs were substantially larger, particularly with mean PK data (RSE: 98.9% for k_{cat}, 107% for K_m). The uncertainty was relatively smaller with all individual PK data (RSE: 59.1% for k_{cat}, 66.4% for K_m). The near-perfect correlation coefficient ($r \approx 1.0$) across both simultaneous optimization cases reflects the lack of identifiability in Michaelis-Menten kinetics when data are insufficient.

Two sets of parameter uncertainty estimates were carried forward for PSA in Step 3:

- Single-parameter: k_{cat} = 10.9 (SE = 0.36) from mean PK data. This represents the simpler case of optimizing one parameter at a time, which is a common approach for PBPK model calibration.
- Two-parameter: k_{cat} = 6.41 (SE = 3.79) and K_m = 1.34 (SE = 0.89) from all individual PK data. Of note, the simultaneous estimates from mean PK data were not and gave extremely large SEs driven by the identifiability limitation; the individual PK data provided more stable, albeit still correlated, estimates.

Table 1. Parameter optimization results from PK-Sim based on the simulated DDI study. Variance (Var) values are based on the output of PK-sim parameter optimization, standard errors (SE) are calculated based on the variance (i.e., square root of the variance).

Data	Optimization	Parameter	Estimate	Var	SE	RSE (%)	Covariance	Correlation
Mean PK	k_{cat} only	k_{cat}	10.9	0.13	0.36	3.30	—	—
	K_m only	K_m	3.31	0.01	0.10	3.02	—	—
	k_{cat} and K_m	k_{cat}	14.76	213.57	14.60	98.9	59.52	0.999
		K_m	3.80	16.61	4.08	107	59.52	0.999
All individual PK	k_{cat} only	k_{cat}	12.10	0.15	0.39	3.22	—	—
	K_m only	K_m	2.95	0.01	0.10	3.39	—	—
	k_{cat} and K_m	k_{cat}	6.41	14.34	3.79	59.1	3.38	0.998
		K_m	1.34	0.80	0.89	66.4	3.38	0.998

Step 3: Probabilistic Sensitivity Analysis (PSA)

Single-parameter PSA

Prior to PSA, a deterministic sensitivity analysis was performed to characterize the relationship between k_{cat} and the DDI metrics (AUC_{inf} ratio and C_{max} ratio) for the midazolam–cimetidine interaction

(*Figure 16*). Over the 95% confidence interval of k_{cat} (10.18 to 11.62), both metric ratios showed a near-linear relationship with k_{cat}. This near linearity implies that PSA output distributions will closely mirror the shape of the k_{cat} input distribution, with minimal distortion from non-linear model behavior.

Consistent with this, the distributions of the AUC_{inf} ratio and the C_{max} ratio from PSA closely replicated the shape of the k_{cat} distribution (*Figure 17*; *Table 2*). Log-normal and normal distributional assumptions for k_{cat} produced similar distributional shapes for k_{cat}, reflecting the small relative uncertainty of k_{cat} in this case (RSE ≈ 3%). It should be noted that no negative k_{cat} values were generated using a normal distribution as a prior, given the small uncertainty. When the uncertainty is relatively large, truncated normal distributions are more appropriate to avoid negative parameter values.

Given that the BHMA framework assumes a log-normal distribution for the geometric mean ratio, the log-normal assumption for k_{cat} was carried forward to Step 4 to ensure methodological consistency. To reduce computation time, PSA outputs were derived from linear regression predictions of the relationship between k_{cat} and the metric ratios within the uncertainty of k_{cat}, which was justified by R^2 ≈ 1.0 for both regressions.

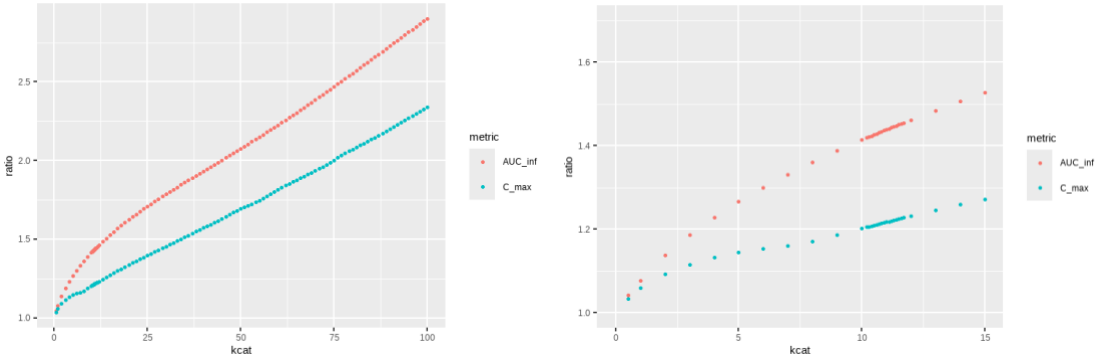
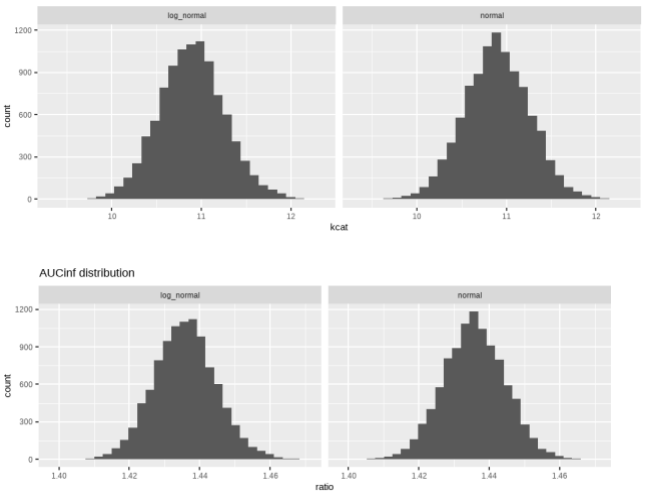


Figure 16. Deterministic sensitivity analysis: relationship between k_{cat} *and the* AUC_{inf} *ratio and* C_{max} *ratio for the midazolam–cimetidine DDI.*



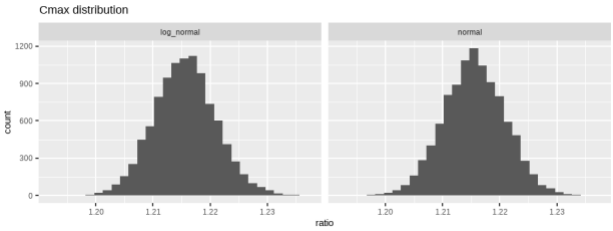


Figure 17. Distributions of k_{cat} *(input) and resultant* AUC_{inf} *ratio and* C_{max} *ratio (outputs) from PSA under log-normal and normal distributional assumptions.*

Table 2. Numerical summary of k_{cat}, AUC_{inf} ratio, and C_{max} ratio from PSA under log-normal and normal distribution assumptions.

Parameter	Distribution	Mean	SE	RSE (%)	Min	Max
k_{cat}	Log-normal	10.902	0.361	3.31	9.639	12.230
	Normal	10.901	0.362	3.32	9.469	12.300
AUC_{inf} ratio	Log-normal	1.436	0.00870	0.61	1.405	1.468
	Normal	1.436	0.00872	0.61	1.401	1.470
C_{max} ratio	Log-normal	1.216	0.00535	0.44	1.197	1.235
	Normal	1.216	0.00536	0.44	1.194	1.236

Two-parameter PSA

PSA was also performed based on the simultaneous two-parameter optimization estimates (k_{cat} and K_m from all individual PK data), using a bivariate log-normal distribution to account for the high parameter correlation (r = 0.998). Due to computation time constraints, 500 simulations were performed.

Compared with the single-parameter case, the input distributions of k_{cat} and K_m were substantially wider and right-skewed, reflecting the higher SEs from simultaneous optimization and the assumption of a log-normal distribution. As a result, the output distributions of AUC_{inf} ratio and C_{max} ratio were also wider and right-skewed, in contrast to the near-symmetric distributions obtained under single-parameter optimization (*Figure 18*; *Table 3*). Nevertheless, the increase in output uncertainty was less pronounced than the increase in input uncertainty, partly because of the near-perfect correlation between k_{cat} and K_m (r = 0.998), which causes their effects on the predicted GMR to partially cancel out.

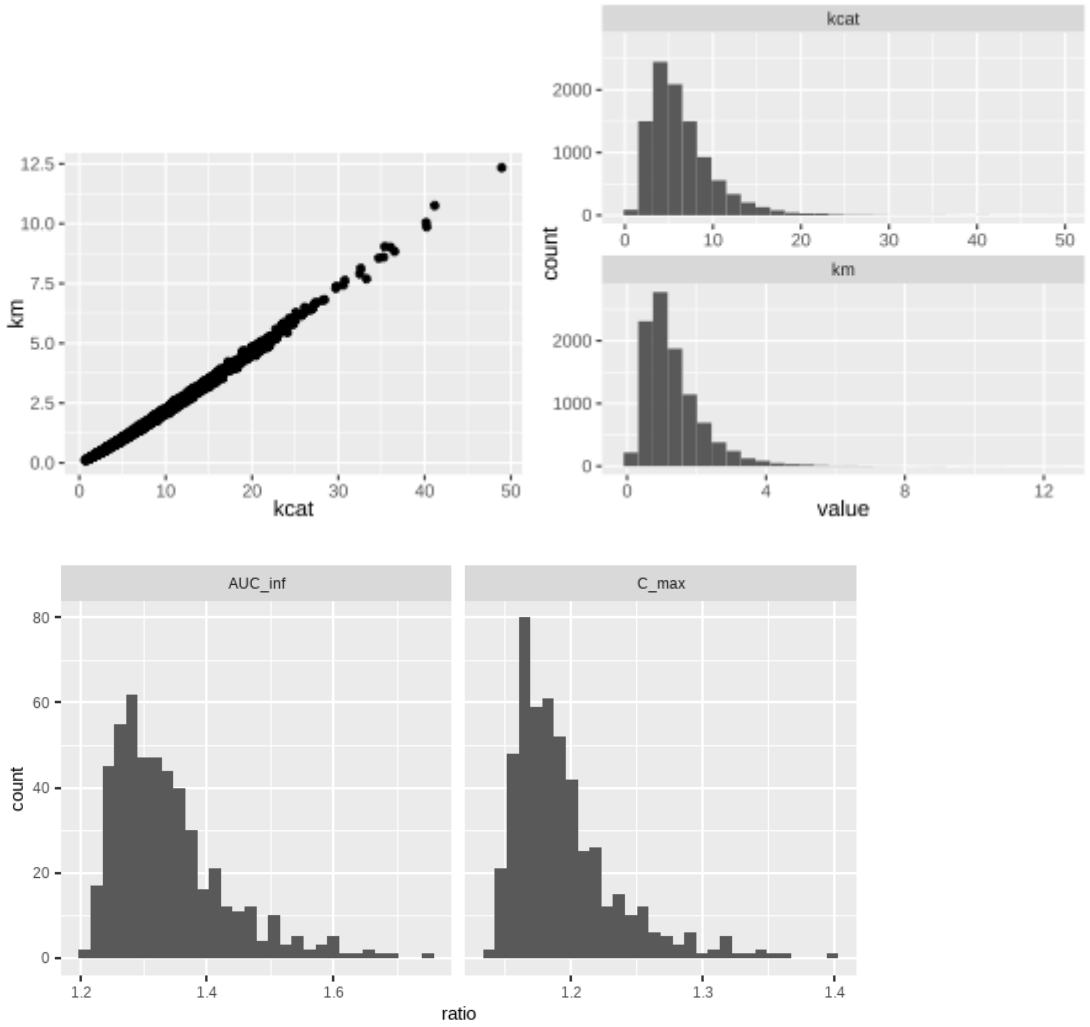


Figure 18. Distributions of k_{cat} *and* K_m *(inputs) and resultant* AUC_{inf} *ratio and* C_{max} *ratio (outputs) from PSA under bivariate log-normal distributional assumptions.*

Table 3. Numerical summary of k_{cat}, K_m, AUC_{inf} ratio, and C_{max} ratio from PSA under bivariate log-normal distribution assumptions.

Parameter / metric	Mean	SE	RSE(%)	Median	Geo. mean	5th pct.	95th pct.
k_{cat}	6.466	3.825	59.16	5.573	5.546	2.232	14.004
K_m	1.353	0.902	66.67	1.129	1.121	0.409	3.123
AUC_{inf} ratio	1.339	0.089	6.65	1.319	1.336	1.239	1.515
C_{max} ratio	1.194	0.040	3.35	1.185	1.194	1.152	1.273

Step 4: Combining PSA and BHMA Uncertainty

The predictive credible interval (PCI) constructed under the four uncertainty schemes for the AUC_{inf} ratio are shown in *Figure 19*.

Within the BHMA-derived intervals, between-study variance was the dominant contributor to interval width: the full BHMA PCI (parameter uncertainty + between-study variance) was substantially wider than the BHMA that excluded between-study variance.

One-parameter PSA

The 90% predictive confidence interval (PCI) from single-parameter PSA (Figure 19 left) was considerably narrower than any BHMA-derived interval. This indicates that the drug-specific parameter uncertainty quantified here is small relative to the platform-level uncertainty captured by BHMA.

Two-parameter PSA

The 90% PCI from two-parameter PSA (Figure 19 right) was wider than that from single-parameter PSA, consistent with the larger parameter uncertainty arising from simultaneous optimization. However, the combined PCI (two-parameter PSA + BHMA) was slightly wider than the BHMA PCI alone.

It should be noted that 500 PSA simulations were performed using PK-sim and were limited by the time-consuming simulation process (4.3 hr.). To match the 10,000 BHMA iterations used in the PCI construction, the 500 simulated AUC_{inf} ratios were replicated 20 times. A future confirmatory PSA with 10,000 independent draws, feasible on a computing cluster, would remove this approximation, though results are unlikely to differ substantially.

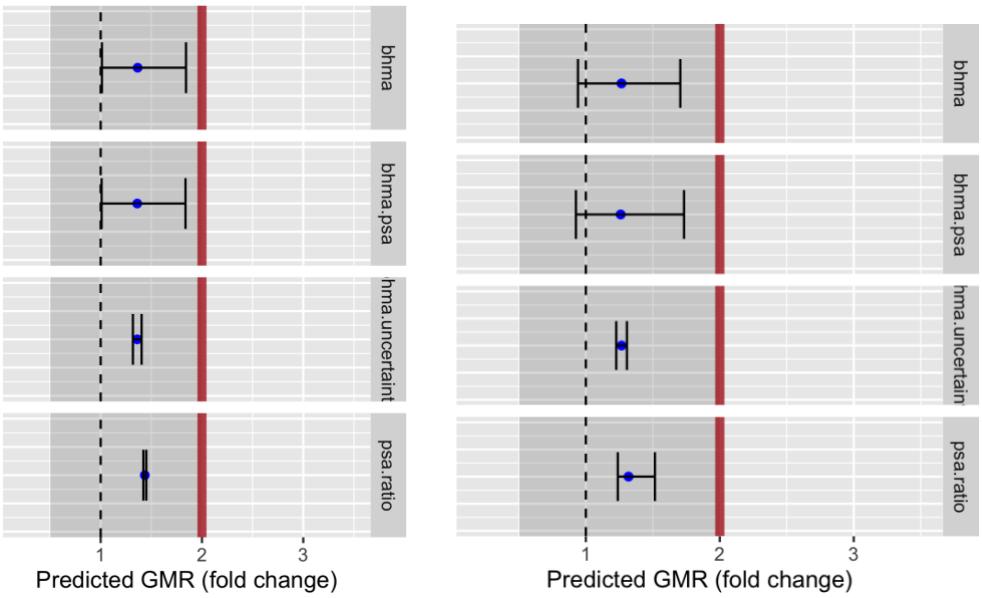


Figure 19. Predictive credible intervals (PCI) for the midazolam–cimetidine AUC_{inf} ratio under four uncertainty combining schemes from one-parameter optimization (left) and two-parameter optimization (right).

5 Discussion

5.1 Part 1: Evaluating BHMA and exploring alternative approach(es) for qualifying PBPK platforms

- **What are the required design characteristics for BHMA?**

- At least 20 DDI studies in the validation dataset are required for BHMA to achieve reliable type I error and power (>90%). The number of subjects per DDI study had negligible impact when ≥ 20 studies were available; it only became relevant with ≤ 10 studies.

- **What are the required operational characteristics of BHMA?**
  - The settings used in the Simcyp qualification opinion (4 chains; 2500 warmup + 2500 sampling iterations) are sufficient for convergence of the Bayesian model, with $\hat{R}$ and ESS meeting target criteria in > 99% of replicates across all scenarios.

- **What is the BHMA performance under different scenarios?**
  - Type I error was controlled across all scenarios regardless of platform bias or between-study variance magnitude. Statistical power was not affected by the bias magnitude but was substantially reduced by high between-study variance in all scenarios. Power increased with the number of DDI studies, suggesting that platform imprecision — rather than systematic bias — is the dominant challenge for qualification.

- **How can the BHMA be used for exposure matching into pediatric populations?**
  - An exposure matching scenario was performed which could be applicable to pediatric dose selection scenarios. To reflect exposure matching scenario, two one-sided tests (TOST) were performed based on exposure limits 0.6 – 1.67 fold difference in GMR (Wang et al). The simulations were based on the same model used for evaluation of DDI scenarios. The results suggested that exposure matching requires more validation datasets compared to the DDI scenario (for CYP inhibition with 2-fold upper limit in GMR). To put this in context of pediatric dose selection, we hypothesize that the number of validation datasets may be more limited for pediatric dose selection compared to DDI predictions for CYP-mediated inhibition scenarios. This indicates that PBPK modeling alone could be too uncertain for pediatric dose selection if the performance is comparable to CYP-mediated inhibition interactions.

    Note that this exposure-matching exercise is exploratory and of limited applicability in the pediatric setting. The adoption of 0.6–1.67 geometric mean ratio bounds and TOST represents a methodological choice rather than broadly accepted criteria. Moreover, adequate exposure matching should address not only central tendency (e.g., mean or median) but also variability (dispersion) across populations, which is not fully captured here. The datasets and BHMA model were developed in a DDI context and, given the limited pediatric data in the public domain and additional uncertainty related to developmental factors, may not adequately reflect the complexities of pediatric extrapolation.

- **Is BHMA robust against model misspecification?**
  - BHMA was largely robust. Non-informative priors (*ninf*) performed equivalently to the base model. The model with informative priors (*inf*) introduced high bias for small dataset sizes but converged with more data. The model without a random effect on BSV (*trad*) showed targeted bias in the $\mu_{\beta_{BSV}}$ parameter and was identified as significantly

different from the base model by ELPD. The lognormal variance model (*LN*) showed very high bias for σstu² at ≤ 50 studies; despite appearing acceptable by ELPD in many replicates, it produced very low type I error, suggesting overly conservative predicted credible intervals.

  - It should also be noted that when a BHMA model is developed for a specific context of use, the impact of relevant covariates should be evaluated. In the context of DDI applications, for example, inhibition type and CYP enzyme should be examined to verify that the bias and/or between-study variance are consistent across covariate levels before excluding them from the model.

- **How was the performance of alternative approach(es) compared to BHMA?**
  - The frequentist approach produced similar bias, RRMSE, and power to BHMA. Parameter estimation with NONMEM performed better as it had a lower bias on the between study variance, σ^2_{stu}. The performance with 50 DDI studies was satisfying if SIR was used to assess parameter uncertainty, but not with only 20 DDI studies. Thus, the frequentist approach without priors is a viable option to the BHMA in the scenarios we investigated here.

5.2 Part 2: PSA for analyzing the uncertainty in model input parameters

- **Is it possible to perform PSA using PK-Sim?**
  - Yes. PSA is feasible using the *ospsuite* R package as a scripting interface to PK-Sim. A practical limitation is computation time: the 2-parameter PSA (500 simulations) took approximately 4.3 hours on a single core. This can be reduced substantially by running on a cluster with multiple cores.

- **What is the impact of the selected prior uncertainty distribution on the PSA result?**
  - The choice of distribution should reflect the available information. When a range of literature values exists, the triangular distribution is appropriate. When a mean and SE are available from parameter optimization, normal or lognormal distributions are more suitable. When the estimated mean is large relative to its uncertainty (as in the k_{cat} single-parameter case: estimate 10.9, SE 0.36), there is little practical difference between normal and lognormal distributions.
  - A more critical driver of PSA uncertainty was the parameter optimization strategy. Simultaneous optimization of k_{cat} and K_m produced near-perfect correlation ($r \approx 1.0$) and very large SEs, particularly with mean PK data, due to structural unidentifiability. This substantially widened the PSA confidence intervals compared with single-parameter optimization, although the resulting uncertainty in the output metric ratios was considerably smaller than that in the input parameters (Table 3), partly due to the partial cancellation of kcat and KM effects on the predicted GMR driven by their near-perfect correlation.

5.3 Part 3: Exploring approaches for combining uncertainty

- **How to understand the uncertainty from PSA and BHMA and its implications for regulatory decision-making?**
  - BHMA uncertainty reflects platform-level prediction performance across historical validation studies. The between-study variance captures: (1) variability in model quality among previously developed PBPK models (including model structure and drug-specific parameters); and (2) inherent between-study variability in DDI outcomes.

- PSA reflects uncertainty in input parameters and their propagation to DDI predictions. It does not capture errors in model structure. The source of parameter uncertainty had a substantial impact on results: single-parameter optimization of k_{cat} yielded small uncertainty in AUC_{inf} and C_{max} ratios (Table 2), whereas two-parameter optimization of k_{cat} and K_m produced relatively wider distribution due to the high parameter uncertainty (Table 3; noted that metric ratio uncertainty was not as high as parameter uncertainty). As a result, the selection of input parameters and source of uncertainty should be well justified for this type of sensitivity analysis.

- **What are the potential approaches to interpret and combine PSA and BHMA uncertainty?**
  - Compare PSA and BHMA uncertainty separately: since BHMA reflects general platform performance, a PSA uncertainty substantially exceeding BHMA uncertainty may indicate that the drug-specific PBPK model is of lower quality than those used in the validation dataset. The interpretation depends on the source of input parameter uncertainty used in the PSA.
  - Add PSA and BHMA uncertainty (reported in this work): drug-specific parameter uncertainty is layered on top of platform uncertainty. This approach may involve double-counting if drug-specific parameter uncertainty is already partially captured in BHMA's between-study variance.
  - Integrate PSA into BHMA estimation (future work): incorporating drug-specific parameter uncertainty directly into the BHMA model structure may avoid possible double-counting and provide a more coherent framework for total uncertainty quantification.

- **What plots are useful to visualize the uncertainty of complex model predictions?**
  - Side-by-side comparison of predicted credible intervals across uncertainty schemes (PSA alone, BHMA, combined PSA and BHMA, and BHMA without between-study variance) allows the relative contribution of each uncertainty source to be assessed visually. Decomposed interval plots highlighting the contribution of between-study variance, parameter uncertainty, and PSA uncertainty would further support regulatory interpretation and decision-making.

Disclaimer

This perspective represents the perspectives of the individual authors and does not represent the views of their institutions.

References

1. Jones H, Rowland-Yeo K. Basic Concepts in Physiologically Based Pharmacokinetic Modeling in Drug Discovery and Development. CPT Pharmacomet Syst Pharmacol. 2013;2(8):63. doi:10.1038/psp.2013.41

2. Kuepfer L, Niederalt C, Wendl T, Schlender J, Willmann S, Lippert J, et al. Applied Concepts in PBPK Modeling: How to Build a PBPK/PD Model. CPT Pharmacomet Syst Pharmacol. 2016 Oct;5(10):516–31. doi:10.1002/psp4.12134 PubMed PMID: 27653238; PubMed Central PMCID: PMC5080648.

3. Tsamandouras N, Rostami-Hodjegan A, Aarons L. Combining the 'bottom up' and 'top down' approaches in pharmacokinetic modelling: fitting PBPK models to observed clinical data. Br J Clin Pharmacol. 2015;79(1):48–55. doi:10.1111/bcp.12234

4. European Medicines Agency (EMA). Guideline on the reporting of physiologically based pharmacokinetic (PBPK) modelling and simulation. 2018. Report EMA/CHMP/458101/2016.

5. European Medicines Agency (EMA). Qualification opinion for Simcyp Simulator. 2025. Report EMADOC-1700519818-2006369.

6. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol. 2014 Dec 1;43(6):1969–85. doi:10.1093/ije/dyu149

7. Stan Docs [Internet]. [cited 2026 May 18]. Stan Reference Manual. Available from: https://mc-stan.org/docs/reference-manual/

8. Gelman A. Two simple examples for understanding posterior p-values whose distributions are far from uniform. Electron J Stat. 2013 Jan 1;7(none). doi:10.1214/13-EJS854

9. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput. 2017 Sep 1;27(5):1413–32. doi:10.1007/s11222-016-9696-4

10. McLatchie Y, Vehtari A. Efficient estimation and correction of selection-induced bias with order statistics. Stat Comput. 2024 Jun 12;34(4):132. doi:10.1007/s11222-024-10442-4

11. Wang Y, Jadhav PR, Lala M, Gobburu JV. Clarification on precision criteria to derive sample size when designing pediatric pharmacokinetic studies. J Clin Pharmacol. 2012 Oct;52(10):1601–6. doi:10.1177/0091270011422812 PubMed PMID: 22162537.

12. Dosne AG, Bergstrand M, Karlsson MO. An automated sampling importance resampling procedure for estimating parameter uncertainty. J Pharmacokinet Pharmacodyn. 2017;44(6):509–20. doi:10.1007/s10928-017-9542-0 PubMed PMID: 28887735.

13. Chen X, Nyberg HB, Donnelly M, Zhao L, Fang L, Karlsson MO, et al. Development and comparison of model-integrated evidence approaches for bioequivalence studies with pharmacokinetic end points. CPT Pharmacomet Syst Pharmacol. 2024 Oct 1;13(10):1734–47. doi:10.1002/psp4.13216

14. Jaynes ET. Information Theory and Statistical Mechanics. Phys Rev. 1957 May 15;106(4):620–30. doi:10.1103/PhysRev.106.620

15. Jaynes ET. Information Theory and Statistical Mechanics. II. Phys Rev. 1957 Oct 15;108(2):171–90. doi:10.1103/PhysRev.108.171

16. Lippert J, Burghaus R, Edginton A, Frechen S, Karlsson M, Kovar A, et al. Open Systems Pharmacology Community—An Open Access, Open Source, Open Science Approach to Modeling and Simulation in Pharmaceutical Sciences. CPT Pharmacomet Syst Pharmacol. 2019 Dec 1;8(12):878–82. doi:10.1002/psp4.12473

17. Hanke N, Frechen S, Moj D, Britz H, Eissing T, Wendl T, et al. PBPK Models for CYP3A4 and P-gp DDI Prediction: A Modeling Network of Rifampicin, Itraconazole, Clarithromycin, Midazolam, Alfentanil, and Digoxin. CPT Pharmacomet Syst Pharmacol. 2018 Oct;7(10):647–59. doi:10.1002/psp4.12343 PubMed PMID: 30091221; PubMed Central PMCID: PMC6202474.

18. Willmann S, Lippert J, Sevestre M, Solodenko J, Fois F, Schmitt W. PK-Sim®: a physiologically based pharmacokinetic 'whole-body' model. BIOSILICO. 2003 Sep 4;1(4):121–4. doi:10.1016/S1478-5382(03)02342-4

19. Open Systems Pharmacology Community. ospsuite: R package to manipulate OSPSuite models [Internet]. 2024. Available from: GitHub. https://github.com/Open-Systems-Pharmacology/OSPSuite-R

20. Olkkola KT, Backman JT, Neuvonen PJ. Midazolam should be avoided in patients receiving the systemic antimycotics ketoconazole or itraconazole. Clin Pharmacol Ther. 1994 May;55(5):481–5. doi:10.1038/clpt.1994.60 PubMed PMID: 8181191.

21. Hanke N, Türk D, Selzer D, Ishiguro N, Ebner T, Wiebe S, et al. A Comprehensive Whole-Body Physiologically Based Pharmacokinetic Drug–Drug–Gene Interaction Model of Metformin and Cimetidine in Healthy Adults and Renally Impaired Individuals. Clin Pharmacokinet. 2020 Nov 1;59(11):1419–31. doi:10.1007/s40262-020-00896-w

22. Fee JPH, Collier PS, Howard PJ, Dundee JW. Cimetidine and ranitidine increase midazolam bioavailability. Clin Pharmacol Ther. 1987 Jan 1;41(1):80–4. doi:10.1038/clpt.1987.13

23. R Core Team. (2025) R A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. - References - Scientific Research Publishing [Internet]. [cited 2026 May 18]. Available from: https://www.scirp.org/reference/referencespapers?referenceid=3967248

24. Sigal MJ, Chalmers RP. Play It Again: Teaching Statistics With Monte Carlo Simulation. J Stat Educ. 2016 Sep 1;24(3):136–56. doi:10.1080/10691898.2016.1246953

25. Gabry J, Češnovar R, Johnson A, Bronder S. cmdstanr: R Interface to 'CmdStan' [manual] [Internet]. 2025. Available from: https://mc-stan.org/cmdstanr/

26. Beal S, Sheiner LB, Boeckmann A, Bauer RJ. NONMEM User's Guides. (1989-2019). Gaithersburg, MD, USA: Icon Development Solutions; 2019.

27. Aoki Y, Nordgren R, Hooker AC. Preconditioning of Nonlinear Mixed Effects Models for Stabilisation of Variance-Covariance Matrix Computations. Aaps J. 2016;18(2):505–18. doi:10.1208/s12248-016-9866-5 PubMed PMID: 26857397.

Appendix A: Tables of results

Table S1: For the core simulations, see csv file: supplementary_UQ_core_simulation.csv

Table S2: For the misspecified models, see csv file: supplementary_UQ_model_misspecification.csv

Table S3: For the exposure matching, see csv file: supplementary_UQ_exposure_matching.csv

Table S4: Base model, real data

Parameter	mean	SD	$\hat{R}$	ESS bulk	ESS tail
ratio_bias	-0.05	0,0173	1	15173	7823
mean_BSV_bias	1,4	0.0933	1	1250	2805
log_sd_BSV_bias	1.25	0.0722	1	1719	3826
extra_var	0.0365	0.00611	1	12488	7751

Table S5: Lognormal priors (N(0,1)), real data

Parameter	mean	SD	$\hat{R}$	ESS bulk	ESS tail
ratio_bias	-0.05	0.0176	1	12201	8030
mean_BSV_bias	1.4	0.0899	1	1219	2176
log_sd_BSV_bias	1.25	0.0715	1	1918	3986

extra_var	0.0389	0.00629	1	10837	7808

Table S6: Lognormal highly informative priors (N(0,0.01)), real data

Parameter	mean	SD	$\hat{R}$	ESS bulk	ESS tail
ratio_bias	-0.0482	0.0168	1	10744	7812
mean_BSV_bias	0.67	0.0841	1	1633	3150
log_sd_BSV_bias	1.47	0.0912	1	1166	2698
extra_var	0.0381	0.00627	1	9730	7569

Table S7: Lognormal noninformative priors (N(0,5)), real data

Parameter	mean	SD	$\hat{R}$	ESS bulk	ESS tail
ratio_bias	-0.0499	0.0176	1	12645	7120
mean_BSV_bias	1.41	0.0923	1.01	1148	2907
log_sd_BSV_bias	1.25	0.0727	1	1926	3714
extra_var	0.0366	0.00615	1	10697	7704

Table S8: Traditional meta-analysis, real data

Parameter	mean	SD	$\hat{R}$	ESS bulk	ESS tail
ratio_bias	-0.0553	0.0186	1	9156	6929
mean_BSV_bias	2.16	0.0304	1	9442	6551
extra_var	0.0307	0.00552	1	9234	6883

Appendix B: Codebase used to conduct the simulation study

The codebase used to conduct the simulation study can be found in the supplementary .zip folder *PBPK_UQ_BHMA-main.zip*. It contains a *readme* file explaining the structure of the codebase.

Part 1: The results of the core simulation study can be found in the *results* subfolder in .csv format. The script used to run the core simulation study, as well as the script used to produce the figures can be found in the *scripts/part1* folder. The results from the frequentist approach without priors, implemented in NONMEM, are also found here. An example control stream for the regular model with $COV is *models/mc-alt_1-1.mod* and an example with SIR is *models/psn.mod.* The script generating the datasets is *scripts/part1/script_generate_nonmem.R*. Since these results were generated through stochastic simulation and estimation (SSE) with 1000 samples, not all files are present but rather some illustrative examples. However, the results of each SSE can be assessed in the summary .csv files located within the corresponding directory (e.g. r*esults/sse_200/raw_results_mc-alt_1-1000.csv*). Processed results from these runs are also found in the *data/t1e_pow.Rds* file. For the runs where SIR was used to get parameter uncertainty, see the *data/t1e_pow_sir_2026-04-29.Rds* file. These .Rds files are generated by *scripts/part1/t1e_pow_SIR.R* and read by the scripts producing graphs *scripts/part1/parse_graph_for_NONMEM_merge_gw_2.R.*

Part 2: The simulated DDI datasets (midazolam and ketoconazole) were generated based on a PBPK model in PKsim. Since the simulations and parameter estimations were carried out using the software PKsim, no code was used in the process. However, the related simulation PKsim files (including the PBPK model and study design setup based on the Olkkolla 1994 study) are provided in *models/PKsim_simulation* as *.pkml* files. The resultant DDI PK datasets can be found in *results/part2* using the format *Olkkolla_[mean/pop]_[control/ddi]_0507.csv*. Those were the resultant datasets after filtered in R (*scripts/part2/generate_DDI_dataset.R*) based on the PKsim simulation output according to the sampling times in the original publication. The rest of operations were performed in R. The related PKsim models (midazolam and cimetidine based on the study from Fee 1987) are in *models/PKsim_simulation*. The R scripts are located in *scripts/part2*. And the results of PSA are in *results/part2*.

Part 3: In the illustration cases, the R scripts for combining the uncertainties of BHMA and PSA are located in *scripts/part3*. The rest of code (DDI dataset simulation and PSA) are listed in Part 2 (see above)

Appendix C: PowerPoint slide deck for dissemination of results

A slide deck with the key findings of the project is available in *CONFIRMS_Deliverable4_UU.pptx*.