ORIGINAL RESEARCH ARTICLE

Towards More Robust Evaluation of the Predictive Performance of Physiologically Based Pharmacokinetic Models: Using Confidence Intervals to Support Use of Model-Informed Dosing in Clinical Care

Marjolein D. van Borselen[1] · Laurens Auke Æmiel Sluijterman[2] · Rick Greupink[1] · Saskia N. de Wildt[1,3]

Abstract

Background and Objective With the rise in the use of physiologically based pharmacokinetic (PBPK) modeling over the past decade, the use of PBPK modeling to underpin drug dosing for off-label use in clinical care has become an attractive option. In order to use PBPK models for high-impact decisions, thorough qualification and validation of the model is essential to gain enough confidence in model performance. Currently, there is no agreed method for model acceptance, while clinicians demand a clear measure of model performance before considering implementing PBPK model-informed dosing. We aim to bridge this gap and propose the use of a confidence interval for the predicted-to-observed geometric mean ratio with predefined boundaries. This approach is similar to currently accepted bioequivalence testing procedures and can aid in improved model credibility and acceptance.

Methods Two different methods to construct a confidence interval are outlined, depending on whether individual observations or aggregate data are available from the clinical comparator data sets. The two testing procedures are demonstrated for an example evaluation of a midazolam PBPK model. In addition, a simulation study is performed to demonstrate the difference between the twofold criterion and our proposed method.

Results Using midazolam adult pharmacokinetic data, we demonstrated that creating a confidence interval yields more robust evaluation of the model than a point estimate, such as the commonly used twofold acceptance criterion. Additionally, we showed that the use of individual predictions can reduce the number of required test subjects. Furthermore, an easy-to-implement software tool was developed and is provided to make our proposed method more accessible.

Conclusions With this method, we aim to provide a tool to further increase confidence in PBPK model performance and facilitate its use for directly informing drug dosing in clinical care.

1 Introduction

Physiologically based pharmacokinetic (PBPK) modeling is an approach for model-informed drug development that has grown significantly over the past decade [1]. Physiologically based pharmacokinetic models describe the absorption, distribution, metabolism, and excretion of a substance in a living system [2]. The models aim to cover the complex processes of drug disposition by integrating anatomical and physiological-related parameters with drug properties

Marjolein D. van Borselen and Laurens Auke Æmiel Sluijterman are joint first authors.

Rick Greupink and Saskia N. de Wildt are joint last authors.

Extended author information available on the last page of the article

to simulate clinical trials [3]. These models have numerous uses in drug development both in industry and clinical practice. The most common purpose of PBPK analyses that are submitted to the regulator is to assess enzyme-based drug–drug interactions, which is followed by the application of PBPK models to estimate pharmacokinetics for special populations [1]. With the rise and refinement of PBPK models, this in silico approach has shown tremendous potential to replace or simplify clinical trials, especially in special populations for which clinical pharmacology data to underpin dosing guidelines are hard to obtain, for example in pediatrics [1, 4]. More specifically, PBPK modeling in pediatrics could be used in clinical care to support off-label dosing, in cases where a drug is prescribed to meet a medical need, but where limited pharmacokinetic (PK) data are available to support a rational dosing regimen.

To use PBPK models for high-impact decisions, thorough qualification and validation of the model is essential to gain enough confidence in model performance [4]. In general, regulatory agencies request that PBPK model performance is assessed against observed outcomes of representative in vivo PK studies [5, 6]. The European Medicines Agency [6] delineates that the acceptance criteria to evaluate the closeness of predicted and observed data depend on the regulatory impact.

However, regulatory agencies do not define how to evaluate model credibility, and also in the scientific literature, an overall consensus is lacking. Kuemmel et al. [7] and Frechen and Rostami-Hodjegan [8] recently proposed the use of a framework to achieve harmonization in the modeling field. One step of this framework introduces the application of context-of-use (COU) and model risk assessment, which determine the required rigorousness of verification and validation activities [7]. Determination of the right validation method is subjective, and there is a lack of consensus on which validation steps should be applied [9]. Furthermore, stakeholders recently concluded a more stringent validation method for model-informed drug development is necessary [10].

Currently, there are numerous validation approaches. It is customary to begin with visual checks. As a second step,

it is common to check if the predicted-to-observed ratios of different PK parameters fall inside a certain *n*fold, twofold being the most common [9]. Other metrics include average fold error, absolute average fold error, and prediction error [11, 12].

These validation steps have limitations and are not thorough enough for high-impact models, such as the direct application of model results for dosing in clinical care. Error metrics (i.e., average fold error, absolute average fold error, and prediction error) can only be applied if the model makes predictions for all separate individuals of the corresponding PK study. This is problematic especially in cases where PBPK modeling is used post-marketing, when often only aggregated data from published PK studies are available in the literature to validate the model. Additionally, the twofold criterion is criticized by many for its wide range, and it is sometimes stated as a criterion only fit for informing drug development or adjusting the dosing of drugs with a broad therapeutic window [9, 13]. Another major limitation of the twofold criterion is that it does not take the inherent randomness of the data into account [13]. The predicted-to-observed ratio depends heavily on the random observed data, especially in small data sets. A good ratio can therefore be a result of mere chance. Although this effect is mitigated somewhat by testing on multiple data sets, it is nevertheless undesirable. This problem was noted by Abduljalil et al. [13] who proposed to test if the predicted geometric mean (GM) falls inside the 99.998% confidence interval (CI) of the true GM. The problem with this approach is that for small data sets, the interval becomes very wide resulting in easy model acceptance while for larger data sets, the model can only be accepted if the predicted and observed GMs are approximately equal, which can never be expected.

We propose a validation step that is more in line with common practices within bioequivalence (BE) studies [14]. From the perspective of using PBPK model predictions for clinical drug dosing, the BE approach appears very useful in assessing PBPK model predictive performance. The methodology is also followed when bringing generic drug preparations to the market and when fulfilling predefined BE criteria, the generic drug is accepted to be equivalent to the brand-name medicine [14]. Along these lines, it would be desirable that PBPK model predictions be measured with similar standards, aside from mechanistical considerations (of course the structure of the model needs to be in order, and parameterizations should have been done in the appropriate mechanistic manner).

For the statistical evaluation, we propose to construct a CI of the predicted-to-observed ratio, and to accept a model if the entire interval falls within predefined boundaries. This is an extension of the twofold criterion that addresses the potential problem of accepting a model merely by chance by now logically incorporating the randomness of the data. In

addition, we suggest to primarily apply BE boundaries [0.8, 1.25] as the predefined boundaries. This 1.25-fold bound has been accepted to test for similarity between test-to-reference drugs and accounts for a 20% clinical variation to determine if the predictive values agree with clinical observed values [14].

For the assessment of BE between drug products, a cross-over design or parallel design can be considered [15, 16]. A cross-over design usually is preferred, as it allows an intra-subject comparison between treatments, as each subject serves as its own control [15]. In a parallel biodesign, each subject is randomly assigned to one of the treatments and both cohorts are compared. Bioequivalence assessment is in this case thus based on both intra-subject and inter-subject variability [15]. Our proposed approach to evaluate PBPK model performance against clinical data uses a similar design, ideally by using individual patient data and simulating the virtual twin, in order to reduce the inter-individual variability comparable to the cross-over design in BE studies. This reduction in variance reduces the required number of clinical observations, as has been shown for BE statistics comparing cross-over and parallel design [15, 16]. However, for post-marketing PBPK studies, often only GM PK parameters are available from the literature, not the individual observations. Therefore, we also provide a group-level approach, comparable with parallel-design BE studies, to create a CI when only aggregated values are given for the clinical data to compare PBPK modeling results. Consequently, as this approach suffers from both intra-individual and inter-individual variance, a larger number of clinical observations is required.

Hence, the aim of this paper is (1) to present our novel approach in more detail, distinguishing two scenarios depending on the availability of individual-level or aggregate clinical data to be used for PBPK model validation, including all formulas; (2) to exemplify the workflow of our approach with published midazolam data from healthy adult populations; and (3) to provide an easy-to-use implementation software tool.

2 Methods

2.1 Description of the Proposed Statistics for the Model Evaluation

#### 2.1.1 General Concept of the Proposed Validation Approach

We propose to test the predictive performance of a PBPK model by evaluating a CI of the predicted-to-observed

geometric mean ratio (GMR) of relevant PK parameters, as intervals are more informative than point estimates, and provide an easy-to-interpret result, as shown in Fig. 1. The CI is constructed for all PK parameters, as requested by regulators, representing relevant parts of the plasma concentration–time curve: area under the curve (AUC), maximum concentration (C_{max}), and half-life ($t_{1/2}$) [6]. Care should be taken that these parameters are obtained from matched simulations to the comparator data set. Model predictive performance is accepted when the CIs of all three PK parameters entirely fall within the predefined boundaries. When one of the CIs entirely falls outside of the predefined boundaries, the model quality is not sufficient and cannot be accepted for its intended use. When the CIs are too wide and do not entirely fall either within or outside the acceptance window, no definitive conclusion can be drawn about the predictive performance of the PBPK model. In those situations, we conclude that, given the amount of variability of the observed data, the data set has an insufficient number of subjects to make a decision. This is similar to BE statistics, where large variability and a smaller number of subjects result in wider 90% CIs [17].

Similarly to BE statistics, our CI approach is not trying to prove that the predictions and observations are identical, but that the size of the difference is not clinically significant [17]. Therefore, defining the significance level of the CI and the boundaries of the acceptance window are adaptable and depend on drug variability and the therapeutic window, as well as the context of use and assessed model risk of the PBPK model of interest [6, 18].

For high-impact applications of PBPK models (e.g., directly informing drug labels, developing clinical dosing guidelines), the exact ranges must be determined by regulatory agencies. We propose a 90% CI and [0.8–1.25] range as boundaries for the AUC and C_{max}, unless the therapeutic window is wide, similar to BE statistics to test for BE [14]. In BE studies, the $t_{1/2}$ is not a parameter of interest, and therefore, the acceptance boundaries of the $t_{1/2}$ are not defined [14]. We propose a 90% CI and [0.67–1.5] range as the boundaries for $t_{1/2}$, as more variability may be present with $t_{1/2}$ estimations. Note that the timing of sampling needs to be accurately matched between clinical measurement and simulation. A complicating factor here is that the exact sampling time of the terminal proportion of the PK curve, i.e., the part that was used to calculate terminal $t_{1/2}$, is often not published with the clinical data. Because of the uncertainty of the terminal sampling points used to calculate $t_{1/2}$ and the difficulty in overcoming the variability associated with this, we propose the broader 1.5-fold acceptance range. We stress that, although we proposed various boundaries, the definitive boundaries should be determined by regulatory agencies.

Fig. 1 Graphical representation of an acceptance window of [0.8–1.25] and the variety of outcomes that are possible following the assessment of a 90% confidence interval around the predicted-to-observed ratio
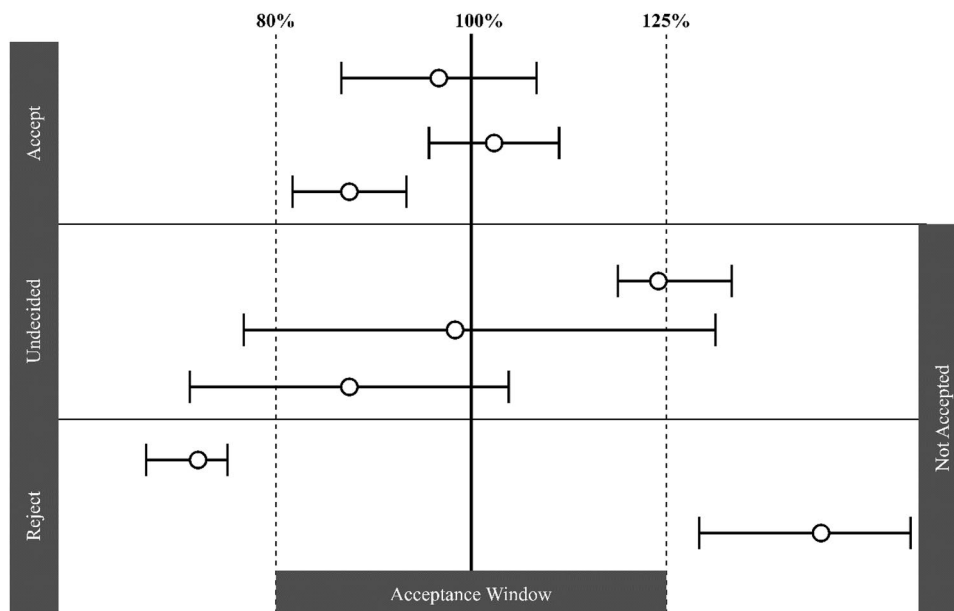
2.1.2 Approach with PK Individual-Level Data

The first approach we propose is based on the cross-over design used in BE studies and uses similar statistics. We believe this is an appropriate method to evaluate the predictive performance of PBPK models with clinical data of individual subjects available for comparison. When the subjects are simulated individually, using as much information about the subject as possible and including it in the modeling effort, a key strength of PBPK modeling, the inter-individual variability reduces, and ideally the resulting difference between observation and prediction will only consist of intra-individual variation. The following steps are proposed to test the model's predictive performance with the individual-level data (I-LD) approach:

1. Decide on the acceptance criterion that meets the COU of the PBPK model. For high-impact models, the BE boundaries of 0.8–1.25 are proposed for AUC and C_{max}.
2. Collect data from clinical studies (e.g., from literature or from investigators) that provide individual PK parameters.
3. Simulate each observed subject separately with the provided characteristics of the subject (e.g., age, height, renal function, where appropriate). A large number ($N = 100$) of simulations per subject are acquired to overcome uncertainty in the predicted PK parameters.
4. Create a 90% CI by using Equation (1) or the online tool, detailed below.
5. Accept, reject, or remain undecided as explained above and illustrated in Fig. 1.

2.1.3 Approach with PK Group-Level Data

Pharmacokinetic I-LD are often not available, as only aggregate data are published and getting access to the original data is often onerous or impossible. Therefore, we also propose a method to evaluate predictive performance with mean values. This group-level data (G-LD) method is comparable to the parallel design sometimes used in BE studies. Similar to the parallel design, the G-LD approach needs correction for both intra-individual and inter-individual variation. As described for BE studies, the intra-individual variability is often smaller (i.e., approximately 15% compared with 30%), which makes the parallel design and G-LD approach subordinate to the cross-over design and I-LD [16]. The greater variability of the G-LD approach results in wider CIs and the necessity of larger data sets to establish predictive performance [17]. The difference with the I-LD approach is that, instead of individuals, now a total population is simulated matching the mean characteristics of individuals enrolled in the clinical study. It is important to note that the GM and not the arithmetic mean should be used both for the observations and predictions, as the AUC, C_{max}, and $t_{1/2}$ are believed to more closely correspond to a log normal distribution [16, 19]. The following procedure of testing model predictive performance with G-LD is proposed:

1. Decide on the acceptance criterion that meets the COU of the PBPK model. For high-impact models, the BE boundaries [0.8 1.25] are proposed for AUC and C_{max}.

2. Simulate a comparable population to the collected clinical study. A very large number (order of magnitude: 10,000) of subjects are simulated in order to remove all uncertainty of the GM estimate. For some specific and very extensive PBPK models and in certain research settings, not enough computational power may be available to complete a large simulation in a reasonable amount of time. The approach is applicable, but the variance of the model, incorporated with the CV of the predictions, will then have an influence on the width of the CI and will need to be taken into account as well.

3. Create a 90% CI by using Equation (2), detailed below, or use the online tool.

4. Accept, reject, or remain undecided as explained above.

2.1.3.1 Construction of the CI using the I-LD approach In this subsection, the statistical background is given for creating a CI of the GMR. More extensive explanation and the mathematical proof of the CI can be found in the Electronic Supplementary Material (ESM).

Suppose there are n observations and predictions, denoted with Y_i and $\widehat{Y}_i$, respectively. The CI is of the following form:

$$\text{CI} = \left[\text{GMR} \cdot e^{-t_{(1-\alpha/2,n-1)}\sqrt{\frac{V}{n}}}, \text{GMR} \cdot e^{+t_{(1-\alpha/2,n-1)}\sqrt{\frac{V}{n}}} \right], \quad (1)$$

Here,

- GMR is the GM of the predictions divided by the GM of the observations, which is equal to:

$$\frac{\left(\prod_{i=1}^{n} \widehat{Y}_i\right)^{\frac{1}{n}}}{\left(\prod_{i=1}^{n} Y_i\right)^{\frac{1}{n}}}.$$

- V is the sample variance of the n differences $\ln\left(\widehat{Y}_i\right) - \ln\left(Y_i\right)$, which is equal to:

$$\frac{1}{n-1} \sum_{i=1}^{n} \left(\left(\frac{1}{n} \sum_{j=1}^{n} \left(\ln\left(\widehat{Y}_j\right) - \ln\left(Y_j\right) \right) \right) - \left(\ln\left(\widehat{Y}_i\right) - \ln\left(Y_i\right) \right) \right)^2.$$

- $t_{(1-\alpha/2,n-1)}$ is the $(1 - \alpha/2)$-quantile of a t-distribution with $n - 1$ degrees of freedom. This value must be looked up in a table or can be calculated using any statistical software package.

2.1.3.2 Construction of the CI using the G-LD approach When only G-LD (GM and geometric coefficient of variation [CV]) are available, the following CI must be used:

$$\text{CI} = \left[\text{GMR} \cdot e^{-t_{(1-\alpha/2,\nu)}\sqrt{V_{\text{obs}}/n_{\text{obs}} + V_{\text{pred}}/n_{\text{pred}}}}, \right.$$

$$\left. \text{GMR} \cdot e^{t_{(1-\alpha/2,\nu)}\sqrt{V_{\text{obs}}/n_{\text{obs}} + V_{\text{pred}}/n_{\text{pred}}}} \right] \quad (2)$$

Here,

- GMR is the observed GMR: $\dfrac{\left(\prod_{j=1}^{n_{\text{pred}}} \widehat{Y}_j\right)^{1/n_{\text{pred}}}}{\left(\prod_{i=1}^{n_{\text{obs}}} Y_i\right)^{1/n_{\text{obs}}}}$.
- n_{obs} is the number of observed subjects.
- n_{pred} is the number of predictions.
- V_{obs} is the sample variance of the observations on a log scale and V_{pred} is the sample variance of the predictions on a log scale. These can be determined from the geometric CV by using $V = \ln\left(\text{CV}^2 + 1\right)$.
- $t_{(1-\alpha/2,\nu)}$ is the $(1 - \alpha/2)$-quantile with ν degrees of freedom, and

$$\nu = \frac{\left(V_{\text{obs}}/n_{\text{obs}} + V_{\text{pred}}/n_{\text{pred}}\right)^2}{\left(V_{\text{obs}}/n_{\text{obs}}\right)^2/(n_{\text{obs}} - 1) + \left(V_{\text{pred}}/n_{\text{pred}}\right)^2/(n_{\text{pred}} - 1)}.$$

Note that if the number of predictions is far greater than the number of observations, this reduces to $n_{\text{obs}} - 1$.

In the limit of a large amount of data or a small variance, the CI converges to the currently used GMR. It can therefore be seen as an easy-to-implement extension of the current testing methodology of the predicted-to-observed ratio.

To aid implementation, we created an online implementation tool. An interactive version can be found at: https://colab.research.google.com/github/LaurensSluyterman/PBPK-evaluation/blob/main/PBPK_evaluation.ipynb. The original GitHub repository can be found at: https://github.com/LaurensSluyterman/PBPK-evaluation.

These formulae also allow us to estimate the required number of subjects for validation. Figure 2 shows the ranges of predicted-to-observed ratios that would lead to acceptance of the model for an increasing number of observations and increasing CV of the observations. We assume that the number of simulations is much larger than the number of observations. In that case, the required number of subjects is determined only by the CV of the observations. The figure suggests that at least 10–15 observed subjects are required if PK I-LD are available, where we expect a lower CV of roughly 30% [20]. At least 25 observed subjects are required if PK G-LD are available, where we expect a much higher CV of roughly 70% [20].
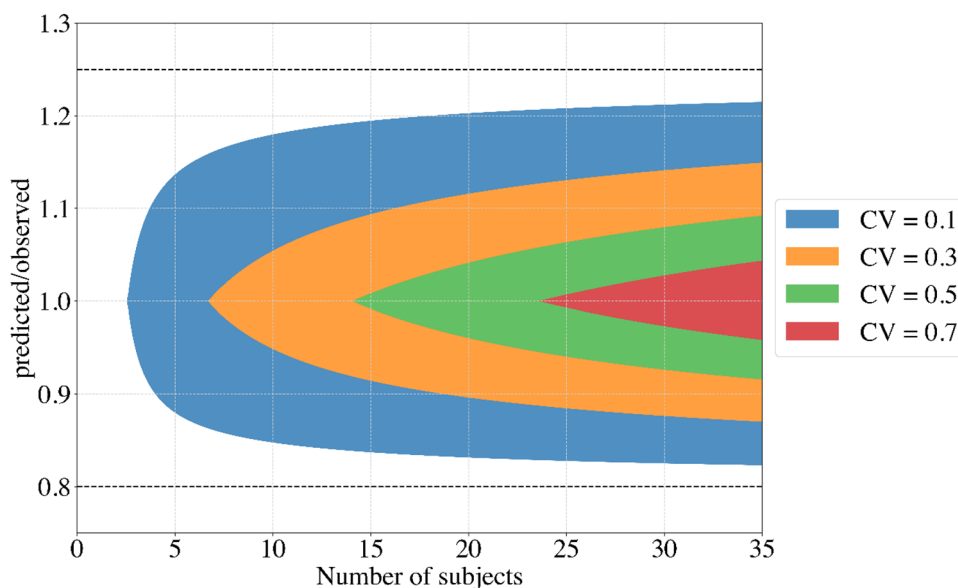
Fig. 2 Illustration of the required number of observations in different scenarios. The shaded areas provide the predicted-to-observed ratios that would lead to acceptance of the model for an increasing number of observations. The different colors correspond to different values of the coefficient of variation (CV). We observe that (1) with minimal variance a data set of five observed values can be sufficient to have a chance to accept the model; (2) more subjects lead to a larger probability of reaching a conclusion; (3) with this approach, a data set with an infinite number of subjects will always accept the model if the geometric mean ratio falls within the interval [0.8–1.25]; and (4) with data sets with a larger variance (e.g., a geometric CV of 0.7), at least 30 subjects are required to have a chance to accept a model

2.2 Simulation Experiment: Comparing the Twofold and G-LD Approaches

Our proposed approach is an extension of the *n*fold method. To demonstrate the differences and similarities, we conducted a simulation study. All simulations were performed in Python. The code is made publicly available and can be found at: https://github.com/LaurensSluyterman/PBPK-evaluation/blob/main/2-fold_example.ipynb.

Hypothetical subjects and predictions were simulated from a lognormal distribution. For each experiment, 10,000 sets of subjects and predictions were simulated, where each individual subject or predictions follows a lognormal distribution:

$$\ln\left(Y_i\right) \sim \mathcal{N}\big(\mu, \ln\left(\mathrm{CV}^2 + 1\right)\big),$$

and

$$\ln\left(\widehat{Y}_j\right) \sim \mathcal{N}\big(\hat{\mu}, \ln\left(\mathrm{CV}^2 + 1\right)\big).$$

Here, Y_i and $\widehat{Y}_j$ represent an observation and prediction of the PK parameter of interest, for example the AUC, and CV is the geometric CV. The same variance was used for both the observations and predictions.

Various scenarios were simulated, considering different true GMRs $\frac{e^{\hat{\mu}}}{e^{\mu}}$, CV values, and numbers of subjects. In every simulation, 10,000 predictions were simulated. For each

scenario, the frequency of model acceptance was determined using the twofold method.

Subsequently, we ran various simulations to demonstrate what the results would have been with our method, while using the G-LD approach. To facilitate a comparison to the twofold method, the [0.5–2] acceptance window was used and not the BE acceptance window, which is used in the other experiments. For various scenarios, we determined how often our approach accepts, rejects, or refrains from deciding.

2.3 Illustration of the Novel Evaluation Approach on an Example PBPK Model

To illustrate the application of the CI of predicted-to-observed GMRs as a validation step, we validated an existing PBPK model for midazolam against published PK data from healthy adults. In the following subsections, the model specification and validation of this example PBPK model are described.

2.3.1 Midazolam Clinical PK Data

Midazolam was chosen because it has been widely investigated as its clearance is considered a probe marker to determine cytochrome P450 3A activity [21]. Additionally, a well-established PBPK model is available. A standardized search query was used to search for healthy adult PK data (detailed in the ESM). Titles and abstracts were screened for healthy volunteer midazolam PK studies. To illustrate

the novel predictive performance criterion, we searched for papers with I-LD and contacted research groups to share raw PK data of published papers with aggregate PK parameters.

For I-LD sets, we only identified two older papers with individual-level PK data [22, 23]. Both papers reported a midazolam PK study with six subjects. We selected Smith et al. [22] to illustrate our novel validation approach, as this paper shared more individual characteristics compared with Heizmann et al. [23]. The characteristics of the subjects of the Smith et al. data set are given in Table S5 of the ESM. In addition, Cleary et al. [24] contributed to our study by testing our I-LD and G-LD approach on their healthy volunteer data receiving an oral single dose of midazolam 2 mg. The patient characteristics and study conditions are described previously [24].

For the illustration of the G-LD approach, we chose two recently published midazolam papers with more than 25 subjects [25, 26]. For an extra illustration of our approach in different scenarios, we created an additional hypothetical clinical data set to illustrate different scenarios of the use of our proposed validation approach, when individual-level PK data are available (ESM).

2.3.2 PBPK Model Specifics

To test the applicability of the novel predictive performance criterion, we used a default PBPK model for midazolam in healthy volunteers (model and compound file Simcyp® Version 22). Simulations were conducted to match the patient and study characteristics of the available clinical data sets. For the I-LD approach, a matched simulation of 100 subjects was conducted per individual observed subject with the custom trial option.

For the G-LD approach, a simulation was conducted matching the data sets in age, female ratio, drug dosing, and dosing route. A virtual trial of 10,000 subjects was simulated. The predictive performance was evaluated with the proposed metric of the G-LD approach.

2.3.3 Model Validation

Before modeling, the acceptance boundaries for model validations were established. The predicted-to-observed 90% CI of the AUC within the 0.8–1.25 range was considered acceptable. The C_{max} of midazolam is known to be highly variable, especially for the intravenous application of midazolam [27, 28]. Therefore, a broader acceptance range of 0.67-fold to 1.5-fold was chosen as the acceptance criterion for C_{max}. As motivated earlier, the same 1.5-fold was chosen as an acceptance criterion for $t_{1/2}$.

For the model evaluation, simulations were conducted to match the data sets of the published papers and the hypothetical clinical study. The hypothetical clinical study was used to illustrate possible scenarios of model evaluation, when more PK IL-D are available (described in the ESM). The predictive performance of the model was validated with the published data sets following the I-LD approach and the G-LD approach. Description of PK parameters are illustrated in Tables S6 and S7 of the ESM. Pharmacokinetic parameters of the cohort of Cleary et al. are described previously [24].

3 Results

3.1 Simulation Experiment: Comparing the Twofold and G-LD Approaches

The following simulation study addresses the difference between the twofold method and our proposed approach. Table 1 illustrates the behavior of the twofold method. Sets of subjects of various sizes were simulated with varying amounts of variance. For each combination, 10,000 sets of subjects and predictions were simulated. The tables provide the acceptance rates for scenarios with a different true GMR, specifically one true GMR within [0.5–2.0], one true GMR just outside the boundaries, and two true GMRs representing mis-specified models. The acceptance rate is defined as the fraction of simulations for which the simulated GMR lies within the twofold range [0.5–2.0].

We observe that the twofold method works as desired when dealing with easier scenarios: a very poor (e.g., a true GMR of 0.3) or very good model (e.g., a true GMR 0.7), small amount of variation in the clinical data (i.e., low CV), and a large number of subjects in the clinical data set. A model with a true GMR, $\frac{e^{\hat{\mu}}}{e^{\mu}}$, of 0.7 that is tested on 20 subjects is accepted in roughly 99–100% of simulations for all CV values. Similarly, a very bad model with a true GMR of 0.3 is nearly always rejected.

Problems arise for the more difficult scenarios where we are testing a model that is slightly below the acceptance threshold, for instance a true GMR of 0.45, on a small data set with a large amount of variance. In these scenarios, a bad model is frequently accepted by the twofold method.

Figure 3 illustrates such a problematic example. The histogram shows 10,000 simulated GMRs, $\frac{\left(\prod_{i=1}^{n_{pred}} \hat{Y}_i\right)^{1/n_{pred}}}{\left(\prod_{i=1}^{n_{obs}} Y_i\right)^{1/n_{obs}}}$, where we used a true GMR, $\frac{e^{\hat{\mu}}}{e^{\mu}}$, of 0.45, five subjects, and a CV of 0.8 for both the predictions and the simulations. For each simulation, 10,000 predictions were simulated. Using the twofold method, the model would have been accepted in 37% of the simulations even though we know that the true GMR, 0.45, falls outside the acceptable window. Because of the variability of the data, the twofold method yields unreliable results for smaller data sets.

Table 1 Acceptance rates using the twofold approach for various scenarios

True GMR 0.7

Number of subjects

		5	10	15	20
	0.8	0.859	0.934	0.966	0.982
	0.7	0.889	0.955	0.979	0.990
CV	0.6	0.911	0.975	0.991	0.996
	0.5	0.944	0.985	0.997	0.999
	0.4	0.973	0.997	1.000	1.000

(a)

True GMR 0.45

Number of subjects

		5	10	15	20
	0.8	0.374	0.317	0.287	0.260
	0.7	0.359	0.305	0.258	0.229
CV	0.6	0.338	0.272	0.237	0.199
	0.5	0.303	0.236	0.197	0.156
	0.4	0.268	0.188	0.149	0.112

(b)

True GMR 0.4

Number of subjects

		5	10	15	20
	0.8	0.235	0.163	0.115	0.082
	0.7	0.213	0.130	0.084	0.058
CV	0.6	0.181	0.100	0.059	0.037
	0.5	0.149	0.068	0.032	0.019
	0.4	0.094	0.032	0.012	0.006

(c)

True GMR 0.3

Number of subjects

		5	10	15	20
	0.8	0.053	0.008	0.002	0.001
	0.7	0.034	0.005	0.001	0.000
CV	0.6	0.021	0.001	0.000	0.000
	0.5	0.008	0.000	0.000	0.000
	0.4	0.002	0.000	0.000	0.000

(d)

CV coefficient of variation, *GMR* geometric mean ratio
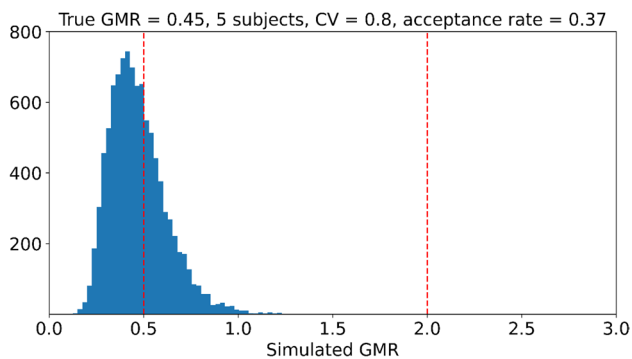

Fig. 3 10,000 realizations of the observed geometric mean ratio (GMR). For each simulation, five subjects and 10,000 predictions were simulated with a coefficient of variation (CV) of 0.8

Table 2 Simulation results of our method for various combinations of CV, *n*, and GMRs

True GMR	CV	*n*	Accept (%)	Reject (%)	Undecided (%)
0.3	0.5	20	0	99.8	0.2
0.3	0.8	20	0	93.9	6.1
0.45	0.5	20	0.5	26.1	73.4
0.45	0.8	20	1.0	17.4	81.6
0.45	0.8	5	3.1	11.1	85.8
0.7	0.5	5	22.2	0.8	77.0
0.7	0.5	20	93.1	0	6.9
0.7	0.8	20	66.8	0	33.2
1.2	0.5	15	99.2	0	0.8
1.2	0.8	30	98.7	0	1.3
1.2	0.8	15	86.3	0	13.7

In scenarios with insufficient information, our model refrains from making a decision

CV coefficient of variation, *GMR* geometric mean ratio, *n* number of subjects

Our extension, not only providing the ratio but a CI of this ratio, addresses this issue. For a fair comparison, we used acceptance boundaries of [0.5–2]: we accept the model if the entire CI falls within [0.5–2], reject if it falls entirely outside this interval, and refrain from deciding if the interval is partially inside [0.5, 2]. Note that for the sake of comparison, we do not use the BE boundaries of [0.8, 1.25] here because that would make a comparison to the twofold approach impossible.

Table 2 gives the results of various simulations. In extreme cases with many subjects or a very low variance, our approach leads to the same conclusions as the twofold method. This is of course to be expected because our method is an extension. For the difficult cases, a GMR close to 0.5 or 2, and a smaller number of subjects, our approach will frequently conclude that there is insufficient evidence to decide and will refrain from doing so.
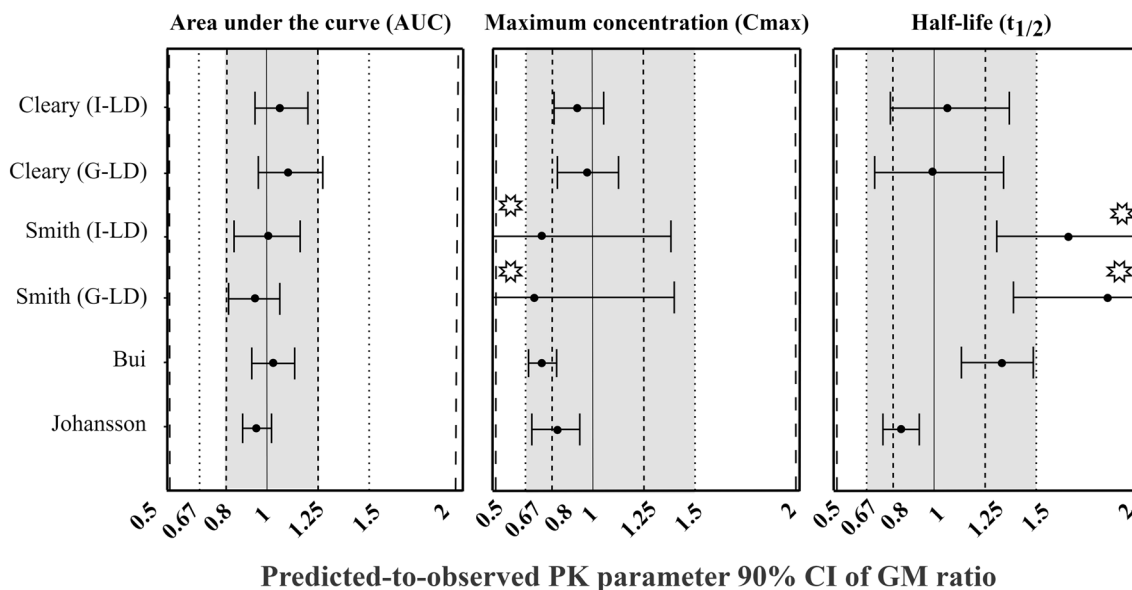
Fig. 4 Predicted-to-observed 90% confidence interval (CI) of geometric mean ratios of the area under the curve (AUC), maximum concentration (C_{max}), and half-life ($t_{1/2}$). For each pharmacokinetic (PK) parameter, we compared the predictions with the published manuscripts with the group-level data (G-LD) approach and for Smith et al. with the individual-level data (I-LD) approach. The acceptance window (gray-shaded area) for AUC was set at the 1.25-fold range, and for C_{max} and $t_{1/2}$ was set at the 1.5-fold range. Maximum concentration and $t_{1/2}$ CIs of Smith et al. fall outside the twofold range and are indicated with a star. *GM* geometric mean

3.2 Validation of the Example PBPK Model

In Fig. 4, a graphical representation of the predicted-to-observed 90% CIs in relation to the different predefined acceptance windows for AUC, C_{max}, and $t_{1/2}$ are shown. At the utmost left and right of the figure, the commonly applied twofold acceptance criterion is also included as dotted vertical lines. The results of the example validation with the clinical data of Cleary et al. illustrate the difference between the I-LD and G-LD approach. Using the I-LD approach, all 90% CIs fall within the acceptance window. With this method, the PBPK model would have been accepted for the study of Cleary et al. Contrary to the I-LD approach, 90% CIs created for Cleary et al. with the G-LD approach are wider. The 90% CI for AUC violates one of the predefined boundaries. Therefore, the conclusion of the predictive performance of the midazolam PBPK model would be to remain undecided.

For the predictions corresponding to the data set published by Smith et al., the 90% CI for AUC falls within the acceptance window. For C_{max} and $t_{1/2}$, the CIs violate one of the predefined 0.67-fold to 1.5-fold boundaries for these parameters. In fact, they would even not have been accepted using the 0.5-fold to 2-fold boundaries. Furthermore, note that when only the GM value of the predicted-to-observed ratio would have been taken into account (ignoring variance, as commonly is done when applying the twofold acceptance criterion), this would have led to accepting the model. In this example validation step, the validation with Smith

et al. would have led to the conclusion to remain undecided about PBPK model predictive performance, as not all CIs of the PK parameters entirely fall within the acceptance window. Yet, we also cannot reject the model, as not one of the CIs falls completely outside the predefined boundaries. The PBPK model would not have been accepted for high-impact applications for populations comparable to those studied by Smith et al. (healthy, male, aged 21–22 years), owing to a lack of observed data.

To give an extra illustration of differences between the I-LD and G-LD approach (Table S4 and Fig. S1 of the ESM), we included results of our evaluation procedure on the hypothetical clinical data set to further illustrate the outcome of the I-LD and G-LD validation approaches should ten instead of six subjects be included in the clinical study or should more extensive details of subject characteristics be available to better match individual predictions. Comparing the approaches, six subjects would be sufficient for the I-LD compared with ten subjects for the G-LD approach to result in CIs for all the PK parameters that fall entirely within the acceptance window.

For the G-LD approach, the PK data of the published manuscripts of Bui et al. and Johansson et al. were compared to the predictions of the example midazolam PBPK model. All 90% CIs of the predicted-to-observed GMRs fall within [0.8–1.25] and [0.67–1.5]. Therefore, the predictive performance is accepted (Table 3).

Table 3 Predicted-to-observed 90% confidence intervals of geometric mean ratios for the different published manuscripts with clinical pharmacokinetic data of midazolam

Clinical data sets	AUC	C_{max}	$t_{1/2}$	Predictive performance
Cleary et al. (N = 27)				
I-LD approach	1.057 [0.933–1.202]	0.933 [0.814–1.072]	1.052 [0.796–1.389]	Accepted
G-LD approach	1.097 [0.942–1.278]	0.979 [0.842–1.139]	0.995 [0.726–1.365]	Undecided[a]
Smith et al. (N = 6)				
I-LD approach	1.006 [0.858–1.179]	0.765 [0.430–1.360]	1.690 [1.307–2.160]	Undecided[a]
G-LD approach	0.936 [0.812–1.079]	0.735 [0.386–1.397]	1.831 [1.395–2.403]	Undecided[a]
Bui et al. (N = 46)				
G-LD approach	1.032 [0.919–1.159]	0.764 [0.694–0.842]	1.318 [1.176–1.477]	Accepted
Johansson et al. (N = 28)				
G-LD approach	0.937 [0.854–1.027]	0.823 [0.725–0.933]	0.841 [0.755–0.936]	Accepted

Geometric mean ratios with 90% confidence interval in []

AUC area under the plasma concentration–time curve, *C_max* maximum concentration, *G-LD* group-level data, *GM* geometric mean, *I-LD* individual-level data, *t_{1/2}* half-life

[a]The model is not validated because the confidence interval does not completely fall within the acceptance window. This is caused by a lack of subjects in the observation data set and too much inter-individual variability

4 Discussion

With this paper, we responded to the call in the PBPK modeling field for a thorough validation approach for high-impact decisions, such as informing clinical drug dosing. We have presented a novel validation approach to evaluate PBPK model predictive performance with a 90% CI of the GMR of PK parameters. The metric is based on BE statistics and is both applicable when individual-level PK data or only mean values of clinical PK studies are available. We believe that a 90% CI of predicted-to-observed GMR is more informative than the currently used ratios. Furthermore, we advocate for a stringent acceptance window of 0.8-fold to 1.25-fold to test if model performance is equivalent to clinical observations.

In our opinion, a metric based on BE statistics is convincing for PBPK model performance evaluation. The goal of our CI approach aligns with BE studies, aimed at demonstrating that the difference between products is not clinically significant [17]. Similarly, the aim of PBPK model validation is not to prove that predictions are identical to a clinical trial, but that the difference is not clinically significant. An advantage of implementing an approach in line with BE studies is that it is a well-investigated field and familiar for stakeholders, regulators, and clinicians. In line with this, our approach for evaluating PBPK model performance mirrors the cross-over and parallel design, depending on the availability of individual-level clinical data.

For the G-LD approach, we advise to use a large number of simulated subjects (N = 10,000) to overcome uncertainty in the GM of the predictions. The prediction of many subjects removes the influence of the variability of PBPK model simulations. This results in a smaller CI. Nevertheless, for

complicated models, simulating large trials may not always be feasible because of insufficient computer processing power. In such cases, the G-LD approach can still be used as it accounts for the variability of the PBPK model predictions within the statistical analysis.

Our study sides with recent critiques of the twofold criterion. van der Heijden et al. [29] argued that more stringent ranges for predicted-to-observed ratios should be considered. Another limitation of the twofold ranges was highlighted by Guest et al. [30] and Abduljalil et al. [13] who demonstrated the necessity of incorporating variability of the PK data into the evaluation procedure. In addition, Guest et al. criticized the broad twofold boundaries and also proposed the use of 1.25-fold boundaries of the BE testing for the predicted-to-observed GMRs. In line with these suggestions, our approach provides a more reliable extension of current validation methodology by adding the use of a 90% CI of the GMR.

We exemplified this reliable extension by comparing our approach to the currently most used twofold criterion [9] . The simulation study, comparing both approaches, showed that in extreme cases both metrics perform identically. However, the twofold criterion proves unreliable when assessing a PBPK model with a GMR close to the acceptance boundaries or when using PK studies with a small number of subjects or high variance. In contrast, we demonstrate that in such difficult scenarios, where evidence is insufficient for a decisive judgment, our approach refrains from doing so.

In the context of the midazolam PBPK model illustration, all GMRs for Smith et al. fell within the twofold range. However, the 90% CIs of the GMRs revealed uncertainty regarding model predictive performance. Our example demonstrated the advantage of CIs over ratios, as ratios alone do

not account for variance and are a less stringent evaluation metric.

Through the example of validating the midazolam PBPK model, we aimed to illustrate the difference between the I-LD and G-LD approaches. Our illustrations, using data of Cleary et al. and a hypothetical trial, clearly demonstrate that the I-LD approach is superior to the GM approach. The I-LD approach requires fewer observations to generate a sufficiently narrow CI for assessing predictive performance. Notably, even though the Cleary et al. data set consisted of 27 subjects, the high variability of the observed AUC and CV of 49% resulted in a slightly too wide 90% CI [24]. This example demonstrates that the I-LD approach, following a cross-over design, requires fewer subjects compared with the G-LD approach, which follows a parallel design.

We mention in our methods the adaptability of the acceptance boundaries. Making decisions on the definitive acceptance window, as is the case with BE studies, will be the role of the regulatory agencies, incorporating details related to the therapeutic window of the drug and the COU of the model. For instance, different boundaries can be justified between applying a PBPK model for drug development or for model-informed dosing recommendations to special populations (i.e., renal disorders or pediatric patients). As BE studies do not assess $t_{1/2}$, we proposed acceptance boundaries of 0.67-fold to 1.5-fold for the $t_{1/2}$. We justify the choice of broader boundaries because of the challenge to mimic the clinical sampling scheme of PK studies with PBPK modeling. The PBPK models, per definition, allow intensive sampling in the terminal phase of the PK curve and use this for calculating $t_{1/2}$, where in many PK studies the terminal $t_{1/2}$ is often based on only two or three sampling points obtained in the terminal phase of drug disposition. Despite our proposal of various boundaries, the regulatory agencies must decide on the definitive acceptance boundaries for the PK parameters.

A limitation to the direct applicability of the method is the lack of openly accessible PK data sets. PK I-LD sets are especially difficult to acquire, but also publications with PK G-LD often are not applicable for a PBPK model evaluation because of lacking the GM and CV or a clear description of the investigated study population. The difficulty of data sharing has also been addressed in a recent workshop of stakeholders as a challenge for further advancement of the PBPK modeling field [10]. We side with their statement of the need for innovative and responsible ways to share I-LD to improve the predictability of PBPK models.

In summary, the strengths of this novel approach compared with the current validation methods are as follows:

- 90% CIs of GMRs are more informative than prediction ratios based on mean values only.

- Our approach is easy to implement with an online tool that quickly generates the intervals after loading the predictions and observations.
- No thorough biostatistical background is required as the approach generates easy-to-interpret results.
- The approach enables the possibility of remaining undecided about model performance. Instead of outright rejecting or accepting the model, the outcome can be that it is impossible to make a decision given the limited amount of random data.
- The approach is applicable as a validation method both with PK I-LD, as well as with only PK G-LD data.
- The validation approach is not specific to any particular PBPK modeling platform and can be applied for results generated by different PBPK platforms than Simcyp®.

5 Conclusions

The applicability of a novel, more stringent validation approach has been demonstrated with an example PBPK model. The approach could be a next step in the harmonization of validation activities in the PBPK field, as the current methods are diverse and the acceptance window is arbitrary. With this method, which has been implemented in an openly accessible online tool, we hope to pave the way to a further increased confidence in PBPK model performance. This, in turn, facilitates its use, not only in drug development but also for directly informing drug dosing in clinical care.

Declarations

Conflict of interest Marjolein D. van Borselen, Laurens Auke Æmiel Sluijterman, Rick Greupink, and Saskia N. de Wildt have no conflicts of interest that are directly relevant to the content of this article. Saskia N. de Wildt reported receiving data safety monitoring board membership fees from Khondrion paid to Radboud University Medical Center. No other disclosures were reported.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material The data analyzed during the current study are available from the corresponding author on reasonable request.

Code availability The code utilized in this article for the software tool is available on GitHub: https://github.com/LaurensSluyterman/PBPK-evaluation.

Authors' contributions All authors contributed to the writing of the manuscript. MB, LS, and RG designed the research. MB and LS performed the research. MB performed the PBPK modeling. LS contributed the analytical tool and statistics.

References

1. Grimstein M, Yang Y, Zhang X, Grillo J, Huang S-M, Zineh I, et al. Physiologically based pharmacokinetic modeling in regulatory science: an update from the US Food and Drug Administration's Office of Clinical Pharmacology. J Pharm Sci. 2019;108(1):21–5. https://doi.org/10.1016/j.xphs.2018.10.033.
2. Gerlowski LE, Jain RK. Physiologically based pharmacokinetic modeling: principles and applications. J Pharm Sci. 1983;72(10):1103–27. https://doi.org/10.1002/jps.2600721003.
3. Verscheijden LFM, Koenderink JB, Johnson TN, de Wildt SN, Russel FGM. Physiologically-based pharmacokinetic models for children: starting to reach maturation? Pharmacol Ther. 2020;211:107541. https://doi.org/10.1016/j.pharmthera.2020.107541.
4. Luzon E, Blake K, Cole S, Nordmark A, Versantvoort C, Berglund EG. Physiologically based pharmacokinetic modeling in regulatory decision-making at the European Medicines Agency. Clin Pharmacol Ther. 2017;102(1):98105. https://doi.org/10.1002/cpt.539.
5. FDA. Guidance for industry: physiologically based pharmacokinetic analyses: format and content. 2018. https://www.fda.gov/media/101469/download. Accessed 17 May 2023.
6. EMA. Guideline on the reporting of physiologically based pharmacokinetic (PBPK) modelling and simulation. 2018. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-reporting-physiologically-based-pharmacokinetic-pbpk-modelling-simulation_en.pdf. Accessed 17 May 2023.
7. Kuemmel C, Yang Y, Zhang X, Florian J, Zhu H, Tegenge M, et al. Consideration of a credibility assessment framework in model-informed drug development: potential application to physiologically-based pharmacokinetic modeling and simulation. CPT Pharmacomet Syst Pharmacol. 2020;9(1):21–8. https://doi.org/10.1002/psp4.12479.
8. Frechen S, Rostami-Hodjegan A. Quality assurance of PBPK modeling platforms and guidance on building, evaluating, verifying and applying PBPK models prudently under the umbrella of qualification: why, when, what, how and by whom? Pharm Res. 2022;39(8):7331–48. https://doi.org/10.1007/s11095-022-03250-w.
9. Sager JE, Yu J, Ragueneau-Majlessi I, Isoherranen N. Physiologically based pharmacokinetic (PBPK) modeling and simulation approaches: a systematic review of published models, applications, and model verification. Drug Metab Dispos. 2015;43(11):1823–37. https://doi.org/10.1124/dmd.115.065920.
10. Jean D, Naik K, Milligan L, Hall S, Huang SM, Isoherranen N, et al. Development of best practices in physiologically based pharmacokinetic modeling to support clinical pharmacology regulatory decision-making: a workshop summary. CPT Pharmacomet Syst Pharmacol. 2021;10(11):1271. https://doi.org/10.1002/psp4.12706.
11. Sheiner LB, Beal SL. Some suggestions for measuring predictive performance. J Pharmacokinet Biopharm. 1981;9(4):503–12. https://doi.org/10.1007/BF01060893.
12. Obach RS, Baxter JG, Liston TE, Silber BM, Jones BC, Macintyre F, et al. The prediction of human pharmacokinetic parameters from preclinical and in vitro metabolism data. J Pharmacol Exp Ther. 1997;283(1):46–58.
13. Abduljalil K, Cain T, Humphries H, Rostami-Hodjegan A. Deciding on success criteria for predictability of pharmacokinetic parameters from in vitro studies: an analysis based on in vivo observations. Drug Metab Dispos. 2014;42(9):1478–84. https://doi.org/10.1124/dmd.114.058099.
14. FDA. Statistical approaches to establishing bioequivalence. 2001. https://www.fda.gov/media/70958/download. Accessed 17 May 2023.
15. Chow S-C, Wang H. On sample size calculation in bioequivalence trials. J Pharmacokinet Pharmacodyn. 2001;28(2):155–69. https://doi.org/10.1023/A:1011503032353.
16. Rani S, Pargal A. Bioequivalence: an overview of statistical concepts. Indian J Pharmacol. 2004;36:209–16.
17. Pearce GA, McLachlan AJ, Ramzan I. Bioequivalence: how, why, and what does it really mean? J Pharm Pract Res. 2004;34(3):195–200. https://doi.org/10.1002/jppr2004343195.
18. Karalis V, Macheras P, Van Peer A, Shah VP. Bioavailability and bioequivalence: focus on physiological factors and variability. Pharm Res. 2008;25(8):1956–62. https://doi.org/10.1007/s11095-008-9645-9.
19. Brett M, Weimann H-J, Cawello W, Zimmermann H, Pabst G, Sierakowski B, et al. Parameters for compartment-free pharmacokinetics standardisation of study design. In: Cawello W, editor. Data analysis and reporting. Aachen, Germany: Shaker Verlag; 1999.
20. Jusko WJ. Perspectives on variability in pharmacokinetics of an oral contraceptive product. Contraception. 2017;95(1):5–9. https://doi.org/10.1016/j.contraception.2016.07.019.
21. Thummel KE, Shen DD, Podoll TD, Kunze KL, Trager WF, Hartwell PS, et al. Use of midazolam as a human cytochrome P450 3A probe: I. In vitro-in vivo correlations in liver transplant patients. J Pharmacol Exp Ther. 1994;271(1):549–56.
22. Smith M, Eadie M, Brophy TOR. The pharmacokinetics of midazolam in man. Eur J Clin Pharmacol. 1981;19:271–8. https://doi.org/10.1007/BF00562804.
23. Heizmann P, Eckert M, Ziegler W. Pharmacokinetics and bioavailability of midazolam in man. Br J Clin Pharmacol. 1983;16(S1):43S-S49. https://doi.org/10.1111/j.1365-2125.1983.tb02270.x.
24. Cleary Y, Gertz M, Grimsey P, Günther A, Heinig K, Ogungbenro K, et al. Model-based drug-drug interaction extrapolation strategy from adults to children: risdiplam in pediatric patients with spinal

muscular atrophy. Clin Pharmacol Ther. 2021;110(6):1547–57. https://doi.org/10.1002/cpt.2384.

25. Johansson S, Rosenbaum DP, Ahlqvist M, Rollison H, Knutsson M, Stefansson B, et al. Effects of tenapanor on cytochrome P450-mediated drug-drug interactions. Clin Pharmacol Drug Dev. 2017;6(5):466–75. https://doi.org/10.1002/cpdd.346.

26. Bui KH, Zhou D, Agbo F, Guo J. Effect of multiple intravenous doses of lanicemine (AZD6765) on the pharmacokinetics of midazolam in healthy subjects. J Clin Pharmacol. 2015;55(9):1024–30. https://doi.org/10.1002/jcph.515.

27. Hohmann N, Kocheise F, Carls A, Burhenne J, Haefeli WE, Mikus G. Midazolam microdose to determine systemic and pre-systemic metabolic CYP3A activity in humans. Br J Clin Pharmacol. 2015;79(2):278–85. https://doi.org/10.1111/bcp.12502.

28. EMA. Guideline on the investigation of drug interactions. 2012. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/07/WC500129606.pdf. Accessed 17 May 2023.

29. van der Heijden JE, Freriksen JJ, de Hoop-Sommen MA, van Bussel LP, Driessen SH, Orlebeke AE, et al. Feasibility of a pragmatic PBPK modeling approach: towards model informed dosing in pediatric clinical care. Clin Pharmacokinet. 2022;61(12):1705–17. https://doi.org/10.1007/s40262-022-01181-8.

30. Guest EJ, Aarons L, Houston JB, Rostami-Hodjegan A, Galetin A. Critique of the two-fold measure of prediction success for ratios: application for the assessment of drug-drug interactions. Drug Metab Dispos. 2011;39(2):170–3. https://doi.org/10.1124/dmd.110.036103.

Authors and Affiliations

Marjolein D. van Borselen[1] · **Laurens Auke Æmiel Sluijterman[2]** · **Rick Greupink[1]** · **Saskia N. de Wildt[1,3]**

✉ Marjolein D. van Borselen
marjolein.vanborselen@radboudumc.nl

[1] Division of Pharmacology and Toxicology, Department of Pharmacy, Radboud University Medical Center, Geert Grooteplein Zuid 10, 6525 GA Nijmegen, The Netherlands

[2] Department of Mathematics, Radboud University Nijmegen, Nijmegen, The Netherlands

[3] Department of Pediatric and Neonatal Intensive Care, Erasmus MC Sophia Children's Hospital, Rotterdam, The Netherlands