Towards uncertainty assessment-based acceptability thresholds for model validation?

Alexander Kulesza (1,2), Camille Massaux (1), Flora Musuamba (1)
(1) Université Namur, Belgium (2) ESQlabs GmbH, Germany

Get the poster *Check in conference app*

UNIVERSITÉ DE NAMUR

Introduction

Physiologically Based Pharmacokinetic (PBPK) models are increasingly recognized for drug-drug interaction (DDI) assessment, with regulatory agencies like EMA and FDA endorsing their use. The recent EMA qualification opinion issued for SimCyp has shown that uncertainty quantification plays a central role in the qualification of mechanistic models, but different model / platforms may require other methods to be employed. Furthermore, for mechanistic models in general, there still lacks consensus best practice on the prospective definition of assessment metrics and goals as a function of the question of interest the modeling platform is supposed to address1,2.

Contact

Alexander Kulesza
Alexander.Kulesza@unamur.be (or Alexander.Kulesza@esqlabs.com)

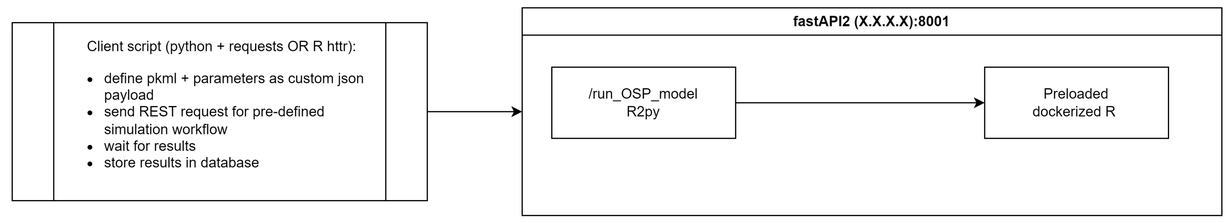
Objectives

The aim of this study is to explore approaches for uncertainty quantification of mechanistic models by numerical experiments and testing metrics and analyses, on the example of a PBPK-DDI use case

EMPIRICAL UNCERTAINTY QUANTIFICATION IN VALIDATION SET ⟺ INPUT ERROR PROPAGATION THROUGH SIMULATION

Methods

We first set up a well-controlled toy model mimicking prototypical DDI, compare assessment metrics and introduce uncertainty in the validation set as an additional dimension to assess and performed bootstrapped statistical tests of detecting relevant DDI in a simulation study and empirically relate other – also frequently used – metrics of goodness-of-fit. We illustrate the essence of this concept on a PBPK model using the Itraconazole-Midazolam DDI as a use case for the original and a range of theoretically weaker interactions. For the PBPK example we used PK-Sim Version 12 Build 369, restoring a published Itraconazole-Midazolam DDI snapshot in the repository GitHub - Open-Systems-Pharmacology/Itraconazole-Midazolam-DDI. With an exported simulation as PKML file and an R script and using the OSP R package we "dockerized" the simulation setup and built a simple API using python fastAPI. Then in a custom python script Ki parameters were modified and time profiles rerun on a Azure cloud server instance and stored in a csv file for (statistical) analysis in a client python notebook using python, numpy, scipy, and sklearn.



Scheme 1: High level view of the cloud-based simulator and client script for generating the Ki – DDI database

Results

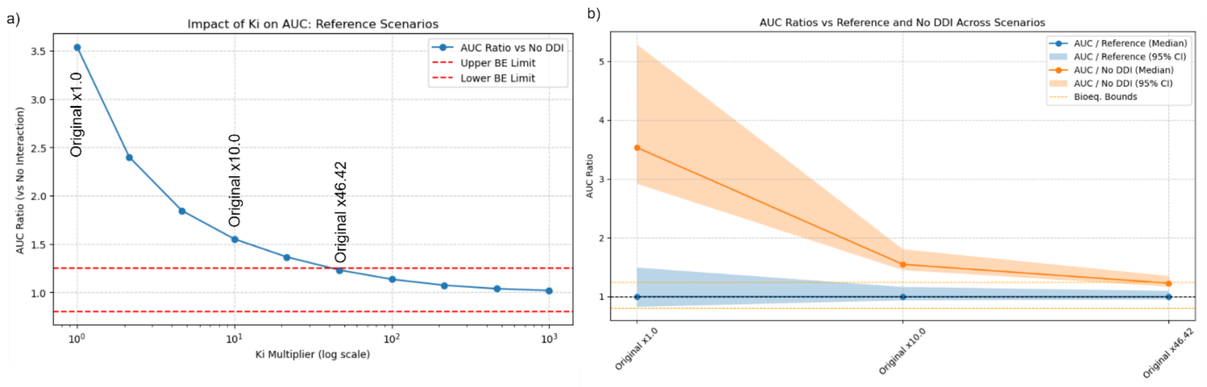


Figure 1: simulated sensitivity of the model to simultaneous scaling of Kis (a) and simulated prediction uncertainty of the AUC ration (between perturbed and no-interaction scenario, orange, as well as reference with error-containing scenarios, blue) assuming a proportional normal error (standard deviation of 0.3, n=200).

Figure 1 shows the sensitivity of the model prediction of the AUC ratio as indicator of DDI interaction strength in vivo, both for the screened grid as well as including assumed Ki uncertainty for 3 selected points. While for the strong inhibitor Itraconazole in the original scenario the uncertainty is highest, misclassification of DDI (e.g. by bioequivalence bounds is unlikely, weaker theoretical interactions' Ki-uncertainty related prediction interval risk to overlap with the latter.

Figure 2 shows a common predicted observed plot of a theoretical validation set (subset of screened values) with 0.5-2 fold and bioequivalence bands. Complementary to that analysis is the actual classification behavior, which is excellent for the selected validation set, but becomes insignificant with less than 15 compounds in the set.

Figure 3 shows a normalized error distribution analysis (Bland-Altman plot) and overlay of limits of agreement with bioequivalence limits to visualize regions (supported by the validation set) that are "safe" or "uncertain" predictions.

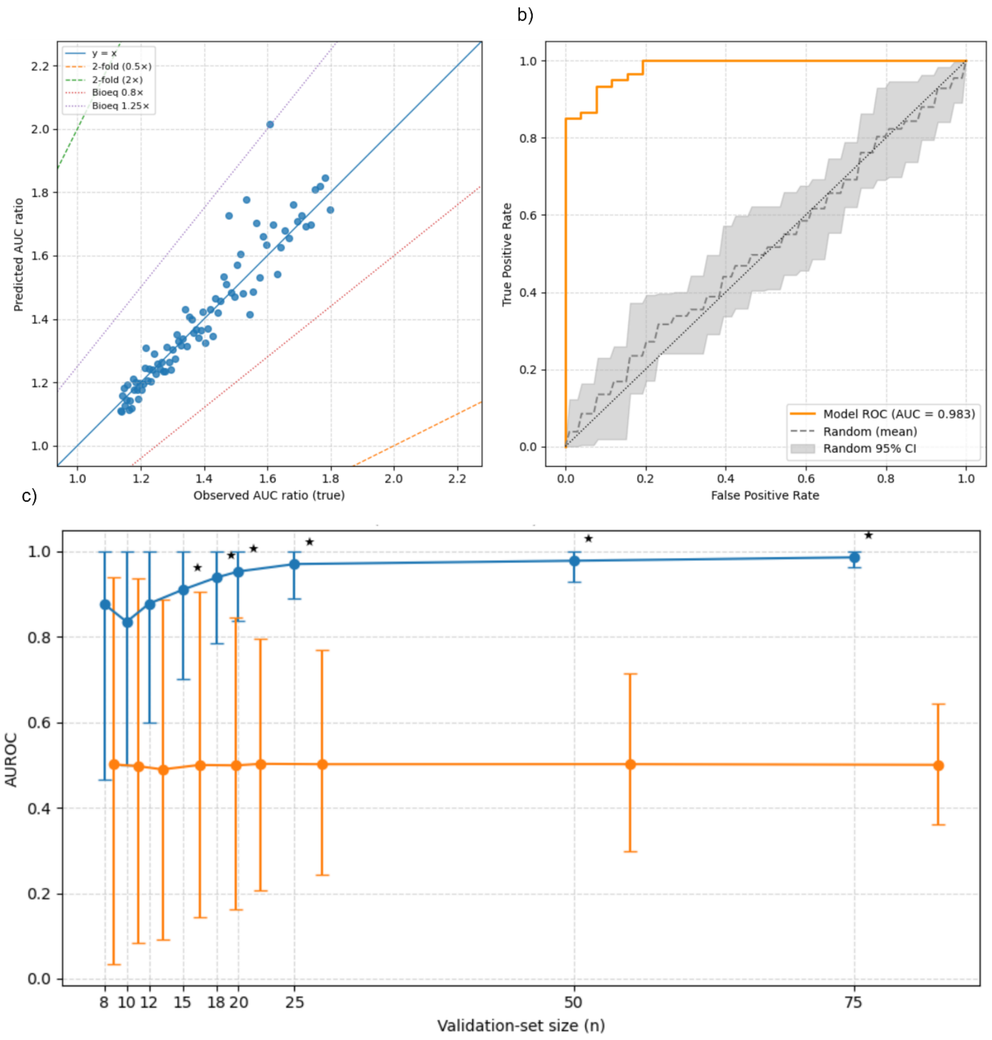


Figure 2: a) "Observed" (reference DDI simulations without assumed error) vs. predicted (DDI simulations with error in Ki, blue points, see main text, an outlier with a predicted AUC ratio > 6 has been omitted) including 0.8-1.25 (dotted lines) and 0.5-2 fold error range (dashed lines); b) ROC analysis for the classification of "DDI" being AUC ratio ≥ 1.25 (orange), including a bootstrapped random permutation (mean ROC in dashed grey line, 95 % confidence interval as grey shaded area). C) model (blue) vs. random classifier (orange) AUROC and 5-95 % CI obtained through bootstrapping and for a series of validation set sizes (x-axis). Stars denote $p < 0.05$ obtained through a bootstrapped hypothesis test.

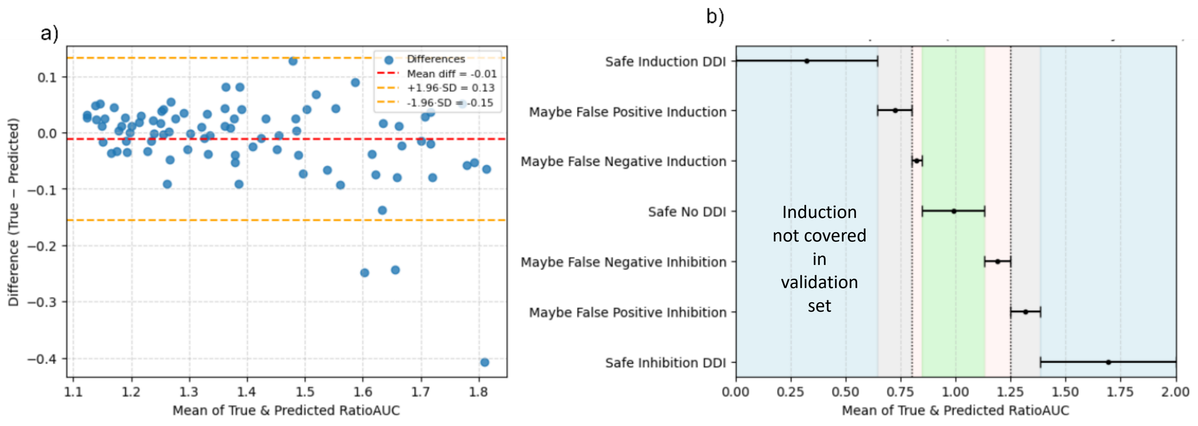


Figure 3: a) Bland–Altman analysis comparing "predicted" and "true" RatioAUC values. The y-axis shows the difference (True − Predicted) against the mean of the two on the x-axis. The red dashed line is the mean difference (bias); orange dashed lines are the limits of agreement (LoA) defined as mean ±1.96 SD. b) Implication for DDI classification. Shaded regions on the x-axis are derived from the bioequivalence thresholds 0.80–1.25 and widened/narrowed by the Bland–Altman LoA. From left to right these zones are: Safe Induction DDI (blue: prediction clearly below 0.80 by more than the lower LoA), Potential False-Positive Induction (grey: prediction <0.80 but within the lower LoA band), Potential False-Negative Induction (pink: 0.80–1.00 but not inside the "safe no DDI" band), Safe No DDI (green, if it exists: interval within 0.80–1.25 after accounting for LoA), Potential False-Negative Inhibition (pink: 1.00–1.25 but not inside the safe band), Potential False-Positive Inhibition (grey: >1.25 but within the upper LoA band), and Safe Inhibition DDI (blue: prediction >1.25 by more than the upper LoA). Horizontal error bars indicate the width of each region to aid interpretation of possible misclassifications.

Conclusion

Uncertainty assessment is an essential element of method qualification and in particular DDI as there are several sources for uncertainty that can be estimated or for which worst case scenarios can be assumed. Sensitivity and error-propagation experiments show how uncertainty in key parameters (for example in vitro Ki) map to AUC ratio prediction errors, which in fact have different risk for misprediction in different DDI regimes. We propose that a PBPK validation and qualification package should: (i) align target decision and a well-established threshold (e.g. bioequivalence) (ii) present error metrics such as predicted-vs-observed error distributions but also overlay those errors with the decision thresholds (iv) evaluate formal statistical tests for essential model performance (e.g better than random classification for which the sample size has a minimal threshold to give a meaningful result) (v) include sensitivity/error-propagation analyses the model performance especially around boundaries where misclassification or misinterpretation can happen.