

## ARTICLE

# Data Digitizing: Accurate and Precise Data Extraction for Quantitative Systems Pharmacology and Physiologically-Based Pharmacokinetic Modeling

Jan-Georg Wojtyniak<sup>1,2</sup>, Hannah Britz<sup>1</sup>, Dominik Selzer<sup>1</sup>, Matthias Schwab<sup>2,3,4</sup> and Thorsten Lehr<sup>1,\*</sup>

In quantitative systems pharmacology (QSP) and physiologically-based pharmacokinetic (PBPK) modeling, data digitizing is a valuable tool to extract numerical information from published data presented as graphs. To quantify their relevance, a literature search revealed a remarkable mean increase of 16% per year in publications citing digitizing software together with QSP or PBPK. Accuracy, precision, confounder influence, and variability were investigated using scaled median symmetric accuracy ( $\zeta$ ), thus finding excellent accuracy (mean  $\zeta = 0.99\%$ ). Although significant, no relevant confounders were found (mean  $\zeta \pm$  SD circles =  $0.69\% \pm 0.68\%$  vs. triangles =  $1.3\% \pm 0.62\%$ ). Analysis of 181 literature peak plasma concentration values revealed a considerable discrepancy between reported and *post hoc* digitized data with 85% having  $\zeta > 5\%$ . Our findings suggest that data digitizing is precise and important. However, because the greatest pitfall comes from pre-existing errors, we recommend always making published data available as raw values.

## Study Highlights

### WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

✓ In quantitative systems pharmacology (QSP) and physiologically-based pharmacokinetic (PBPK) modeling, data digitizer becomes a valuable tool to translate literature data from a graphical representation into numerical values.

### WHAT QUESTION DID THIS STUDY ADDRESS?

✓ This study investigated the usage of digitizing software in QSP and PBPK modeling. Moreover, it evaluated the software accuracy, precision, confounder influence, and variability between software. In addition, the discrepancies between reported and graphically presented data were analyzed.

### WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

✓ The results of this study contributed to an improved understanding of the precision and accuracy of digitizing software.

### HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?

✓ The study findings could help improve the quality of the QSP and PBPK models, which were developed based on digitized literature data. Furthermore, they can protect the modeler from using biased data that could subsequently lead to false *in silico* predictions and hence hamper the drug discovery and development process or, even worse, harm patients as a result of erroneously derived therapy recommendations.

During the past few years, quantitative systems pharmacology (QSP) and especially physiologically-based pharmacokinetics modeling (PBPK) have proven to be an important cornerstone of model-informed drug discovery and development.<sup>1</sup> However, for model development, time-dependent data of pharmacological relevant processes are a crucial requirement. Unfortunately, published data are typically presented in aggregate form as plots or graphs without providing access to the underlying raw, uncondensed data. As a result, researchers must extract the information of interest from the graphical representation to use the data for their modeling approaches. Despite the potential to automatically data-mine population average

pharmacokinetic (PK) data for certain applications,<sup>2</sup> data extraction from graphical representations still requires manual efforts. To illustrate the scale of this issue, it should be noted that PBPK projects not uncommonly rely on extracted data gathered from up to 50 articles.<sup>3–9</sup> Fortunately, several off-the-shelf digitization software packages that allow the extraction of numerical information from their two-dimensional graphical representation are currently available.<sup>10–13</sup> These software solutions have been in active use for some time for the well-established population PK approaches.<sup>14</sup> However, neither for them nor for QSP or PBPK modeling is information available regarding the importance and use of digitizing software. Moreover, to the best of our knowledge,

<sup>1</sup>Clinical Pharmacy, Saarland University, Saarbrücken, Germany; <sup>2</sup>Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany; <sup>3</sup>Department of Clinical Pharmacology, Pharmacy and Biochemistry, University Tübingen, Tübingen, Germany; <sup>4</sup>Cluster of Excellence iFIT (EXC2180) "Image-guided and Functionally Instructed Tumor Therapies", University of Tübingen, Tübingen, Germany. \*Correspondence: Thorsten Lehr (thorsten.lehr@mx.uni-saarland.de)

Received: October 10, 2019; accepted: April 9, 2020. doi:10.1002/psp4.12511

there is no systematic evaluation of the accuracy and precision of these software solutions, nor have any interfering factors that could potentially bias the digitized output been identified. In addition, little is known about the extent of discrepancy between reported and graphically presented data that is typically only revealed after *post hoc* digitization and the nature of these errors and confounding factors when it comes to the digitization process. Consequently, these factors can potentially interfere with model development and evaluation processes and ultimately lead to false predictions and questionable model-based decisions.

Thus, the first objective of this analysis was to assess the general usage of the data digitizing software for QSP and PBPK modeling. Second, this analysis aimed to evaluate the accuracy and precision of a relevant digitizing software. Moreover, discrepancies between reported and graphically presented data were quantified, and the covariates influencing the digitizing process were identified. Finally, recommendations regarding the creation of digitizable plots and graphs and the digitization process itself were proposed.

## METHODS

### Literature search

In a first step, literature published between January 2005 and September 2019 were queried for terms regarding the most common digitizing software in combination with the two terminologies “systems pharmacology” and “physiologically based pharmacokinetic.” For this purpose, the software Publish or Perish<sup>15</sup> was used using the Google Scholar search engine for terminology queries. Google Scholar was used because the search engine allows a full-text analysis, which was a prerequisite because the use of digitizing software is, in most cases, only mentioned in the Methods sections of texts. For each digitizing software, the search query “systems pharmacology OR physiologically based pharmacokinetic AND software name” was used. The search was performed by one author (J.-G.W.) and reviewed by two other authors (H.B., D.S.). The search query results were not further edited or restricted by specific exclusion criteria. Moreover, no gray literature analysis was performed. Subsequently, the annual publications that contain the terms were used as a surrogate marker of importance. Furthermore, Poisson regression was applied to describe and predict the trends in software usage from 2005 to 2021. Moreover, a detailed search for publications from 2012 to 2019 in *CPT: Pharmacometrics & Systems Pharmacology (CPT:PSP)* was also performed. For this, the search query “systems pharmacology OR physiologically based pharmacokinetic AND software name” was used. To identify unreported uses of digitization, all publications from *CPT:PSP* containing only the terms “systems pharmacology OR physiologically based pharmacokinetic” without the name of a digitization software were selected. Following this, the cumulative publication frequency was calculated. Afterward, based on the assumption that data digitization is necessary for every project related to QSP or PBPK, the relative frequency of unreported digitizing software usage was estimated. Finally, to investigate the reporting rate of digitizing techniques in the methods, all published articles from *CPT:PSP* from 2018 were reviewed manually to identify articles related to PBPK that referenced a digitizing software and most likely had used literature data.

### Software evaluation

A study was performed to assess the accuracy and precision of the digitizing software GetData Graph Digitizer<sup>10</sup> and to identify the interfering factors that potentially have an influence on the digitized output. As study inclusion criteria, the subjects had to be at least 18 years old and be able to use the digitization software independently after a standardized software introduction. Furthermore, they had to give informed consent and, following this, were randomly split into two groups (group 1 and group 2). All subjects had to fill out a standardized questionnaire to collect demographics such as age and education. They were asked to digitize the same steady-state plasma concentration-time graph of a hypothetical drug (two-compartment model, absorption constant ( $K_a$ ) = 3 hour<sup>-1</sup>, plasma clearance (CL) = 4 L/hour, central volume of distribution ( $V_1$ ) = 20 L, intercompartmental clearance (Q) = 3 L/hour, peripheral volume of distribution ( $V_2$ ) = 70 L, oral bioavailability ( $F$ ) = 100%, dose = 1 mg simulated using Berkeley Madonna V 8.3.18 developed by Robert Macey and George Oster, University of California, Berkeley, CA) three times in a row. The two-compartment model was chosen as it can be easily parametrized and because the simulations can be easily reproduced. To minimize a possible bias attributed to learning effects, the plasma concentration-time profile in group 1 consists of 10 values marked as circles following 10 values marked as triangles (sample time points: 0, 1, 2, 3, 5, 7, 10, 13, 16, 20, 24, 25, 26, 27, 29, 31, 34, 37, 40, 44). The profile in group 2 was designed as triangles first and circles last. The random allocation sequence for the study subjects was generated with Excel 2016 (Microsoft, Redmond, WA) using a two-sized block randomization and implemented via consecutive questionnaire numbers. No blinding was performed. To validate that demographics are equally distributed within the groups, an analysis of variance (ANOVA) or chi-square goodness-of-fit tests were performed for continuous and categorical demographics, respectively. If the study data were missing, it was imputed with calculated median values for continuous variables and with calculated modal values for categorical variables. To evaluate a potential bias attributed to missing values, statistical analyses were performed with and without imputed values whenever necessary, and the results were compared afterward.

For comparing accuracy and precision, the scaled median symmetric accuracy ( $\zeta$ )<sup>16</sup> and  $\zeta$  standard deviation (SD) were calculated as shown in Eqs. 1–4. Scaling (minimum–maximum normalization<sup>17</sup>) as depicted in Eq. 1 was independently performed for time and concentration values.  $\zeta$  was calculated over the combined vector of scaled values for time and concentration values.

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} + \varepsilon_{\text{machine}} \quad (1)$$

$$Q_i = \frac{X_{\text{scaled, digitized}}}{X_{\text{scaled, simulated}}} \quad (2)$$

$$\zeta = 100 * \left( \exp \left( \text{median} \left( \left| \log_e (Q_i) \right| \right) \right) - 1 \right) \quad (3)$$

$$SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (4)$$

With  $x_{\text{scaled}}$  = scaled value,  $x$  = original value,  $x_{\text{min}}$  = minimum of the original values,  $x_{\text{max}}$  = maximum of the original values,  $\epsilon_{\text{machine}}$  = machine epsilon,  $Q_i$  = accuracy ratio,  $x_{\text{scaled,digitized}}$  = scaled digitized value,  $x_{\text{scaled,simulated}}$  = scaled simulated value,  $n$  = number of  $\zeta$  values, and  $x_i = \zeta$ ,  $\bar{x}$  = mean  $\zeta$ .

Because  $\zeta$  for values equal zero is not defined, the machine epsilon ( $2.2\text{E-}16$ ) was added to each value. Afterward, the impact of demographic variables (age, sex, education, average computer usage per day, experience with digitizing software, and right-handedness) and study-specific variables (digitizing time, symbol shape, type of computer used, and mouse usage) on  $\zeta$  were investigated using multiple linear regression. Moreover, to analyze if multiple digitization leads to better results, an ANOVA with different repetitions on  $\zeta$  was performed.

The required study sample size was calculated for comparing two sample means with an equal variance. Because no literature reference values were available, a mean  $\zeta$  of 5% and a variance of 2% for circle shapes were assumed. Furthermore, a 1.5% increase of  $\zeta$  in the triangle group was assumed compared with circles. From this, the necessity of at least 62 subjects was calculated to get a significant result with a statistical power of 80%, and a significance level of 5% and a dropout rate of 10% were assumed ( $28 \times 2 \times 1.1 = 61.6$ ).

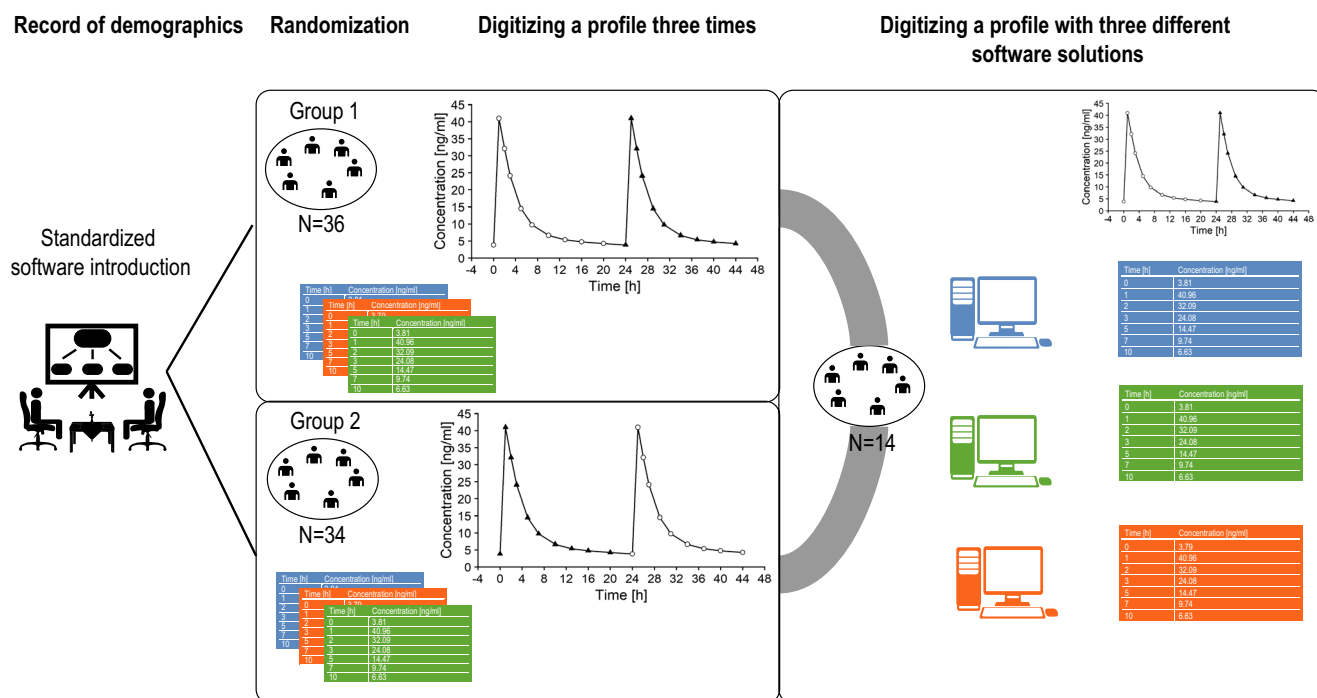
Subsequently, to investigate the impact of the use of digitized data on parameter estimation, the PK parameters of the hypothetical drug were estimated for each of the digitized profiles via nonlinear optimization using the `lbfgsb3` R package.<sup>18</sup> Afterward, the relative deviation compared

to the parameters used for simulation were calculated and visualized.

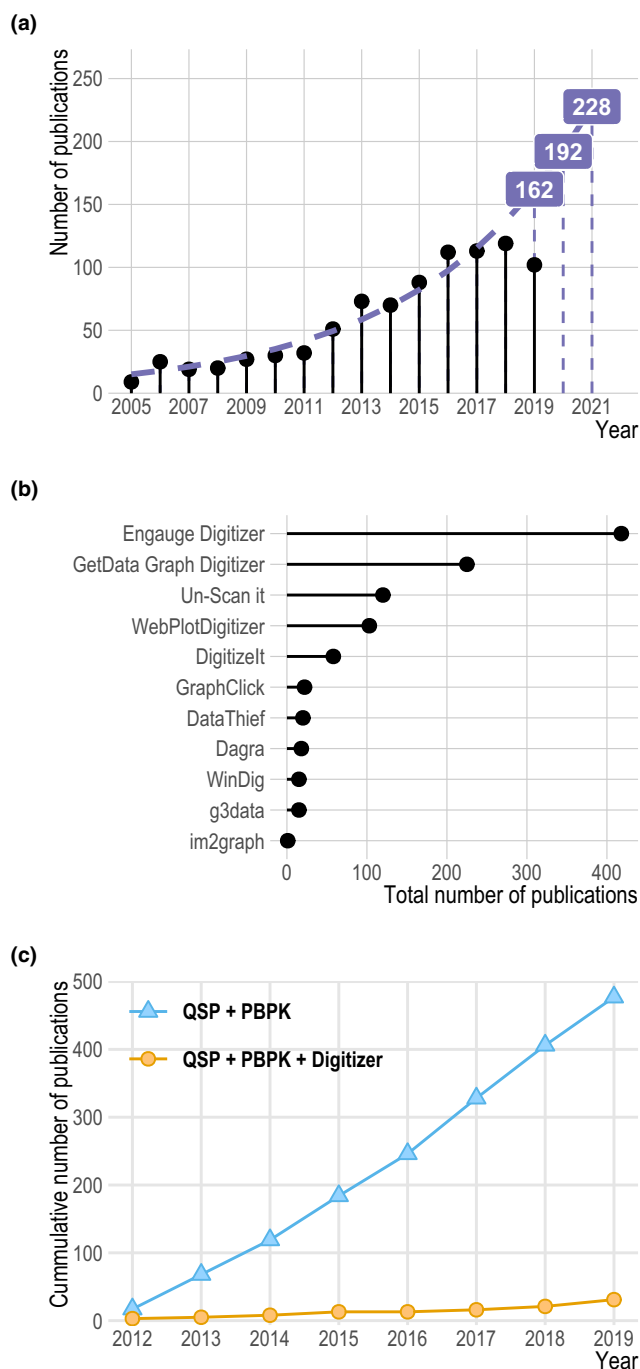
Consequently, a substudy with 14 subjects from the main study group was conducted to compare the accuracy and precision of the digitization software packages DataThief, Engauge Digitizer, and GetData Graph Digitizer. In this study, the subjects digitized the graph from group 1 with each digitization software. Afterward,  $\zeta$  and  $\zeta$  SD were calculated for the digitized profiles and subsequently analyzed via an ANOVA and pairwise *t*-test. The graphs of both groups and an overview of the whole study procedure are shown in **Figure 1**. The three digitization software packages DataThief, Engauge Digitizer, and GetData Graph Digitizer were selected based on the criteria of software availability, usability, included features, and the feedback from a small user survey (10 subjects from our group). A comprehensive list of the different digitization software features can be found in **Table S1**.

### Analysis of published PK data

Finally, the extent of discrepancy between the reported and graphically presented data were investigated based on published sample time points and mean peak plasma concentration ( $C_{\text{max}}$ ) values. For this, digitized readouts as well as published raw values from single and multiple dose profiles that were previously digitized in-house with GetData Graph Digitizer were used. A complete list from all studies included can be found in the supplementary material in Tables S2 and S3. Unscaled  $\zeta$  was calculated individually for all available values. Following this, a stepwise multivariate linear regression analysis (forward inclusion  $P \leq 0.05$ , backward elimination  $P \leq 0.01$ ) was performed to investigate the relationship between the  $\zeta$  values and the portable



**Figure 1** Schematic overview of the study protocol as well as the concentration-time profiles digitized by study subjects.



**Figure 2** Results from literature search. **(a)** Number of publications containing the terms “systems pharmacology” or “physiologically based pharmacokinetic” and the names of the digitization software packages investigated during the past few years. Labels and the dashed purple line shows model-estimated values. Solid lollipops represent the observed values. **(b)** Total number of publications containing the terms “systems pharmacology” or “physiologically based pharmacokinetic” and the name of the digitization software package for each package investigated. The names investigated were “Engauge Digitizer,” “GetData Graph Digitizer,” “Un-Scan it,” “WebPlotDigitizer,” “Digitizelt,” “GraphClick,” “DataThief,” “Dagra,” “WinDig,” “g3data,” and “im2graph.” **(c)** Analysis of the cumulative publication frequency in *CPT: Pharmacometrics & Systems Pharmacology* of the terms “systems pharmacology” or “physiologically based pharmacokinetic” (blue solid line, triangles) when compared with “systems pharmacology” or “physiologically based pharmacokinetic” and the names of the digitization software packages (orange solid line, circles) investigated during the the past few years. QSP, quantitative systems pharmacology; PBPK, physiologically-based pharmacokinetics.

Engauge Digitizer was most frequently cited, followed by the GetData Graph Digitizer (see **Figure b2**). For most software, an increase in use during the analyzed time span was observed. Arithmetic mean increase per year was 16%, with the highest increase for WebPlotDigitizer (87%) and the highest decrease for DataThief (–8%) (see **Figure S1**).

The Publish or Perish search query for QSP or PBPK in the journal *CPT:PSP* from 2012 to 2019 revealed 477 publications. In contrast, for the search terms QSP or PBPK and the names of the most important digitization software packages, only 31 entries were found. Based on the assumption that every QSP or PBPK project requires the use of digitization software, these findings led to an underreporting rate of 94%. **Figure 2c** shows the cumulative number of publications for both search queries. In addition, after the manual review of articles published in *CPT:PSP* in 2018, 20 original research articles that presented PBPK modeling results were found. Among those, 16 used concentration-time or other PK data for model development or validation. Among the 16 studies, 12 used literature data that most likely had to be digitized, 3 had access to individual level data, and 1 study used data from a database for their model development. Among the 12 studies that used literature information, 17% ( $n = 2$ ) reported that they had used a digitization software, leading to an underreporting rate of 83%. A detailed overview of the manual review process is shown in **Figure S2** and **Table S4**.

### Software evaluation

Overall, 70 subjects (51% male) were enrolled in our study. Their mean age was 30 years (range of 18–65 years), and they engaged in a mean computer usage of 4.1 ( $\pm 3.0$ ) hours per day. Only 4% of them had worked with a digitizing software before. All subjects had an educational degree, with the lowest being lower secondary school–leaving certificate and the highest being a doctorate. Demographic characteristics and the number of subjects were equally distributed in both groups as summarized in **Table 1**. The ANOVA and chi-square goodness-of-fit tests revealed no significant differences in study demographics between the groups, with

document format (PDF) (scanned vs. not scanned), the publishing year, and the investigated parameter ( $C_{\max}$  or sample time points). In addition, for values that revealed a  $\zeta$  greater than 5%, a root cause analysis was performed.

All graphical and statistical analysis as well as the sample size calculations were performed using R and R-Studio.<sup>19,20</sup>

## RESULTS

### Literature search

Digitizing software is increasingly used in QSP and PBPK modeling as shown in **Figure 2a**. The free and open-source

**Table 1 Study demographics**

Parameter and descriptive measures	Group 1, N = 36	Group 2, N = 34	Total, N = 70
Age, y, mean (SD)	30 (13)	30 (12)	30 (13)
Average computer usage per day, hour, mean (SD)	4.5 (3.1)	3.7 (2.9)	4.1 (3.0)
Mouse usage, count (%)	31 (86)	30 (88)	61 (87)
No experience with digitization software, count (%)	33 (92)	34 (100)	67 (96)
Right-handedness, count (%)	30 (83)	32 (94)	62 (89)
Male, count (%)	20 (56)	16 (47)	36 (51)

SD, standard deviation.

*P* values always greater than 0.08. The demographic information for all study participants was complete. Thus, no data had to be imputed.

Mean  $\zeta$  was small for all digitized profiles (0.99%), indicating excellent accuracy. Furthermore,  $\zeta$  SD was low (0.72%), revealing a good precision of the software. The regression analysis revealed a significant ( $P < 0.001$ ) effect of symbol shape and digitizing time on  $\zeta$ . These effects are visualized in **Figure 3**. Triangles had a 1.9-fold increased mean  $\zeta$  when compared with circles (1.3% vs. 0.69%) and, hence, were less accurately digitized. Furthermore, subjects digitizing slowly were more accurate than subjects digitizing faster (**Figure 3b**). Besides that, no other covariates had a significant effect on  $\zeta$ .

From the first to the last repetition, the mean digitizing time declined moderately (first, 3.01 minutes; second, 2.24 minutes; third, 2.16 minutes). No statistical difference in accuracy or precision was observed between the three replicates as shown in **Figure S3**.

Furthermore, the estimated PK parameters based on the digitized profiles revealed only small deviations when compared with the parameters used for profile simulation with a mean modulus of the relative deviation of 0.5%. The deviation of all parameters is visualized in **Figure 3c**.

An ANOVA analysis of the performed substudy revealed statistically significant differences in accuracy and precision among the investigated software (**Figure 3d**). GetData Graph Digitizer and Engauge Digitizer had a similar mean  $\zeta$  value (0.2%), whereas DataThief had a markedly increased value (0.5%). The  $\zeta$  SD was 0.1%, 0.2%, and 0.4% for GetData Graph Digitizer, Engauge Digitizer, and DataThief, respectively.

### Analysis of published PK data

For investigating the literature profile quality, 181 mean  $C_{max}$  values and 3499 sample time points of concentration-time profiles obtained from 81 literature studies published between 1984 and 2017, which were presented as graphs and as numeric values, were analyzed. Digitization was carried out as part of two of our in-house model developments (simvastatin, ketoconazole). The digitized profiles were originally derived from graphs that had either linear or logarithmic scaled axes and were depicted either as single or panel plots. Therefore, 3% of the sample time points and 85% of the mean  $C_{max}$  values had a  $\zeta$  greater than 5%. The linear regression analysis revealed that besides the parameter investigated (sample time points or  $C_{max}$ ), neither the PDF format (scanned vs. not scanned) nor the publishing year had a significant effect on  $\zeta$ .

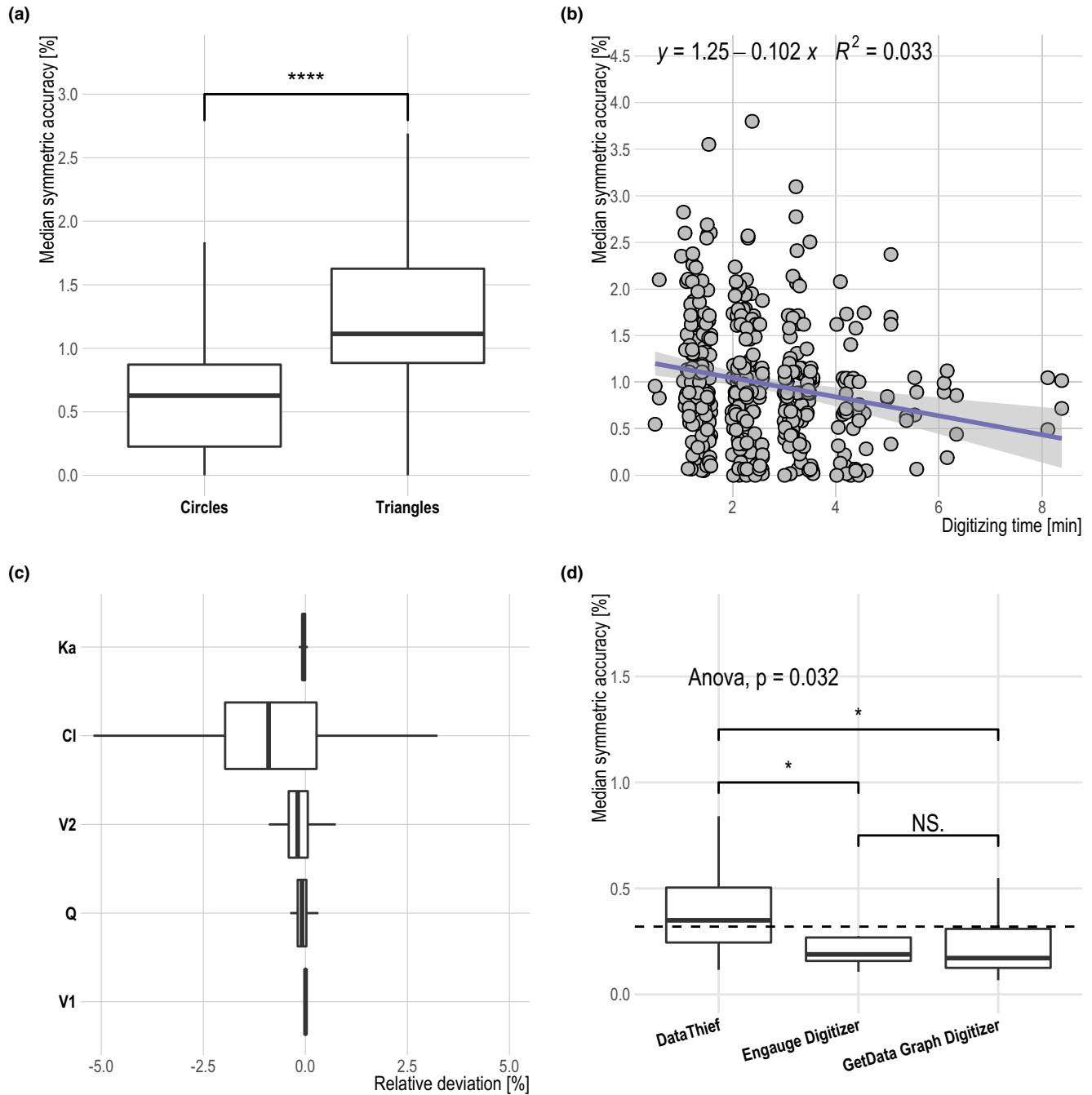
The subsequently performed root cause analysis found for all sample time points with  $\zeta$  greater 5% a justification, namely, either the x axis was not sufficiently resolved or the x axis in the graphic had an uneven resolution. In contrast, a reason for the discrepancy could be identified in only 40% of the mean  $C_{max}$  values with  $\zeta$  greater than 5%. Specifically, they were caused either by poor graphic quality, incorrect labeling, or different types of central tendencies presented in the table and graphic. For the remaining digitized mean  $C_{max}$  values, no justification could be found, leading to an assumption of either incorrectly stated mean  $C_{max}$  values in the depicted concentration-time profile or in the presented table. An overview of the error frequencies and  $\zeta$  distribution is presented in **Figure 4**.

Finally, based on the most important study findings, a digitization algorithm as depicted in **Figure 5** was formulated that can help guide scientists through the digitization process.

## DISCUSSION

### Literature search

The reuse of data through digitization from published articles is an easy-to-use and attractive way for gathering necessary information, especially in QSP and PBPK modeling. This is also evident in the investigated publication frequencies of “systems pharmacology” or “physiologically based pharmacokinetic” in combination with the names of the investigated digitization software solutions. Thus, a remarkable, constant, and exponential increase in the number of literature references was observed. This was observed not only for the pooled number of publication frequencies but also for most of the software packages themselves. However, it should be mentioned that because of the large number of different software solutions, it is very unlikely that all digitizing software available was investigated. In addition, we assume that the actual number of unreported digitizing software usage is significantly higher and that the software is often not reported. This is supported by the cumulative number of publications from 2012 to 2019 in the journal *CPT:PSP*, where 477 publications citing “systems pharmacology” or “physiologically based pharmacokinetic” are published but only 31 publications additionally mention the name of a digitizer software. Subsequently, even if not every publication on the subject requires digitization software, this still suggests a massive underreporting. This assumption is further validated by the manual review of publications

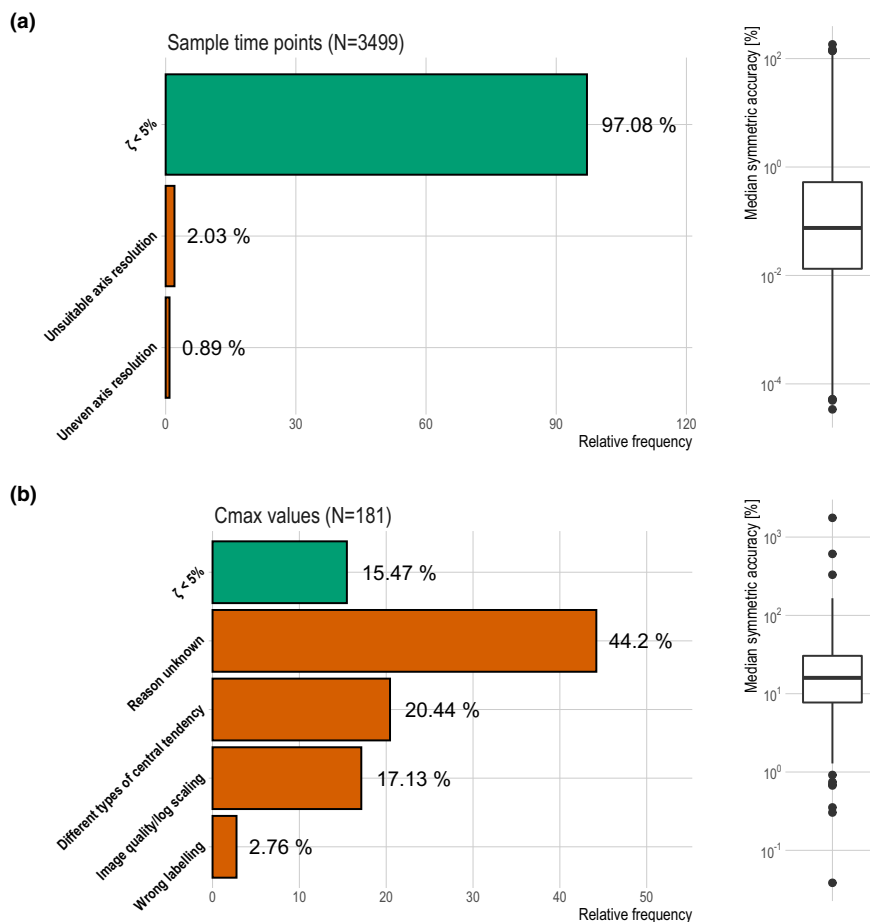


**Figure 3** Results from the multiple linear regression analysis. Influence of the symbol shape (a) on median symmetric accuracy is visualized as a boxplot with the Wilcoxon rank sum test. Effect of digitizing time (b) on median symmetric accuracy is depicted as a scatterplot with linear regression formula and coefficient of determination. (c) Relative pharmacokinetic parameter deviation of the estimated parameters when compared with the values used for profile simulation is shown as boxplots. Parameter estimation was performed for each digitization run. (d) Median symmetric accuracy for different digitization software shown as boxplots. ANOVA as well as pairwise Wilcoxon rank sum tests were performed. ANOVA  $P$  value is stated. For Wilcoxon rank sum test  $P$  values, the following annotation was used: \*\*\*\* $\leq 0.0001$ , \*\* $\leq 0.01$ ; \* $\leq 0.05$ ; NS  $> 0.05$ . In addition, in d arithmetic, the mean of all groups is shown as a dashed line. All boxplots visualize the following descriptive statistics: The median value, the interquartile range, and the 1.5-fold interquartile range. ANOVA, analysis of variance; NS, not significant.

from 2018 in *CPT:PSP* on PBPK modeling, revealing a reporting rate of 17%.

As a drawback of the performed literature search, one can state that the search query results were not further revised, and no gray literature analysis was carried out as

recommended for systematic reviews. However, the purpose of the literature search was not the development of a systematic and exhaustive review but, rather, the identification of general trends. For this reason, the methodology differs from that of a systematic review. For example, search



**Figure 4** Discrepancy between reported and graphically presented sample time points and mean  $C_{max}$  values. Relative frequency of  $\zeta < 5\%$  and justifications for  $\zeta$  values  $\geq 5\%$  were presented as bar charts. Distribution of  $\zeta$  values were in addition shown as boxplots. **(a)** depicts the results for digitized sample time points while **(b)** displays the digitized mean  $C_{max}$  values.  $C_{max}$ , peak plasma concentration.

queries were not carried out in various databases such as PubMed or Embase as recommended for systematic reviews. Instead, the search engine Google Scholar was used, whose algorithm screens many different databases and sources; moreover, this is better suited to get an overview of the frequency of use in literature.

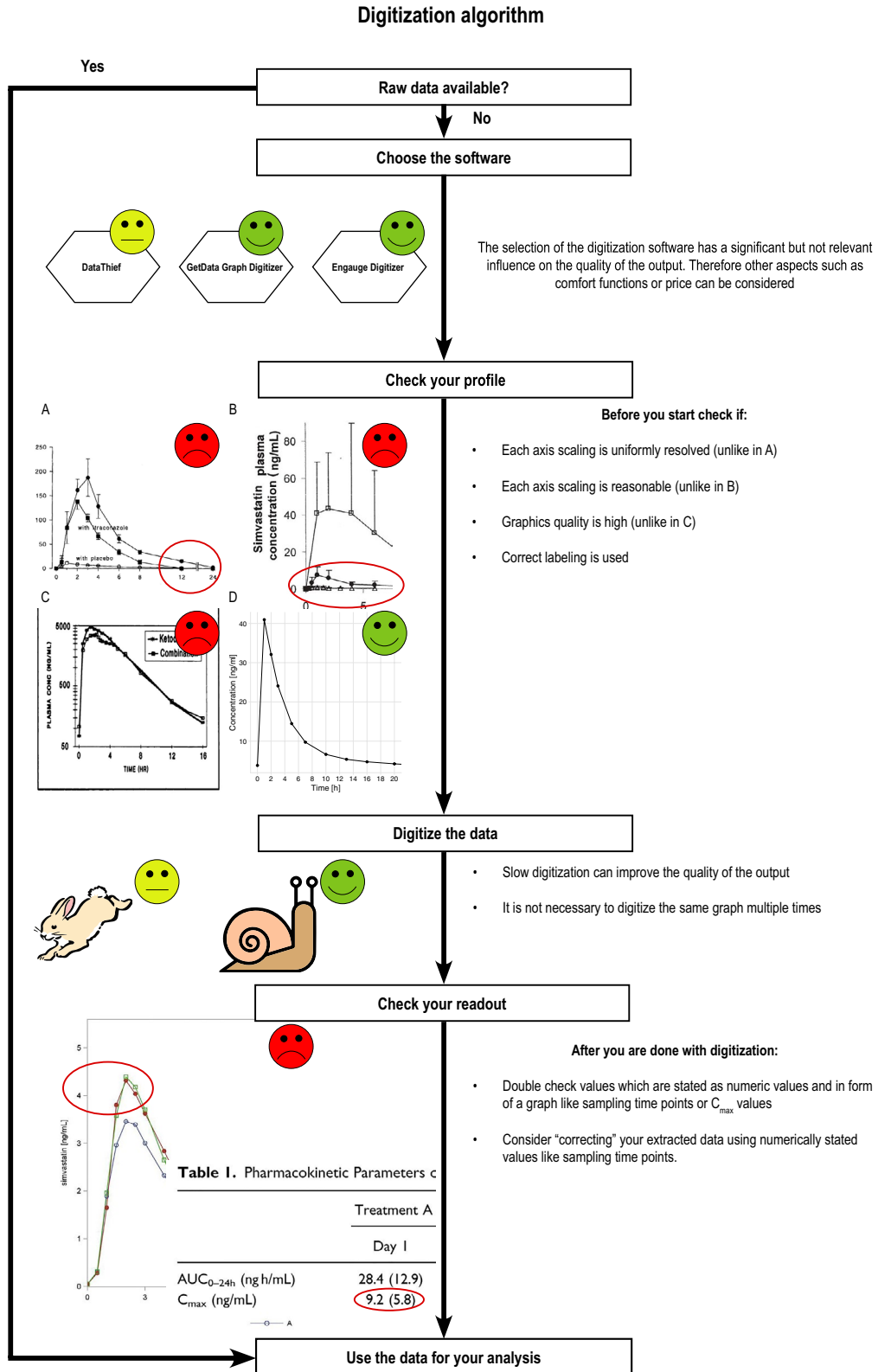
Nevertheless, it can be assumed that the usage of digitizing software in QSP and PBPK modeling will further increase in the next few years as shown in **Figure 2**.

### Software evaluation

For the study with 70 subjects,  $\zeta$  was chosen as an error metric because of its intuitive interpretation as a relative error.<sup>16</sup> With a mean  $\zeta$  of 0.99% and a mean  $\zeta$  SD of 0.725%, the digitized readouts were tremendously accurate and precise. This is also reflected in the accuracy of the PK parameters subsequently fitted to the digitized profiles showing a mean modulus of relative deviation of 0.5% when compared with the original values. This further suggests the assumption that the accuracy of individual digitized values is less important because they are not independently analyzed in the model but as time-dependent series of values. Apart from this, the symbol shape and digitizing time revealed a significant effect on the accuracy ( $\zeta$ ), leading, for example, to a 1.9-fold

lower mean  $\zeta$  value for circles when compared with triangles. Nevertheless, these effects are still negligible, considering the overall small  $\zeta$ . Although the average digitizing time declined with each repetition, no advantages in accuracy were observed if the same graph was digitized more than once. Based on these results, we recommend that one-time, slow-paced digitizing is sufficient for a proper readout.

The additionally performed substudy revealed significant differences in accuracy comparing DataThief and the two software products Engauge Digitizer and GetData Graph Digitizer. Here, DataThief showed a 1.5-fold decline in accuracy when compared with Engauge Digitizer and GetData Graph Digitizer. As mentioned previously, this effect is still negligible because the mean  $\zeta$  for all three software packages was still less than 0.6%. This led to the assumption that although significant differences between the software exist, accuracy is still excellent, and thus, other software features are more important. For example, the freely available and open-source software Engauge Digitizer is still under active development on GitHub, providing a wide range of functionalities and available in different languages for multiple operating systems. Although this might raise the question of whether, apart from the software, other factors such as the operating system also have an impact on



**Figure 5** Proposed digitization algorithm to improve the daily digitizing and graph creation practice in the fields of quantitative systems pharmacology and physiologically-based pharmacokinetics. Examples are taken from refs. 22 to 25.  $AUC_{0-24h}$ , area under the plasma concentration-time curve calculated from 0 to 24 hours post dose;  $C_{max}$ , peak plasma concentration.



accuracy, at least for Engauge, this is very unlikely because it is programmed in Qt, an operating system-independent programming language.<sup>12,21</sup>

### Analysis of published PK data

Data that were redundantly presented as numeric values as well as in a graphs or plots were analyzed using  $\zeta$  as an error metric for the differences between reported values and the corresponding digitized graphical representation. If  $\zeta$  was  $> 5\%$ , the graphs and plots were further explored to determine the article properties that may impede researchers from retrieving correct readouts.  $\zeta$  of the digitized sample time points were in good agreement with the results derived from the previously conducted study. However, a few sources of errors could be identified. Specifically, the resolution of the axes seemed to have an important influence on the quality of the digitized readout. If one of the axis resolutions is uneven or the resolution does not allow cleanly distinguishing between individual measuring points, the result can be falsified. Surprisingly, with 80% of the 181 mean  $C_{\max}$  values having a  $\zeta > 5\%$  and a maximum  $\zeta$  of 1760%, alarmingly large differences between the published numerical values and the values in graphs were found. Even worse, as shown in **Figure 4**, after the performed root cause analysis for 40% of the  $C_{\max}$  values with  $\zeta$  greater 5%, no justification could still be identified. This leads to the assumption that either the wrong graph was plotted or a wrong  $C_{\max}$  was reported. Based on these findings, we strongly recommend that published data should additionally always be made available as raw data. Furthermore, if such access is available, digitizing reported and graphically presented data should be avoided; instead, raw data should be used. Moreover, if access to raw data is not available, researchers should check that each axis scaling is uniformly and optimally resolved, the graphics quality is high, and the correct labeling is used. In addition, they should try to double check their digitized values based on values that are additionally published in a numeric form. However, although following the last recommendation may prevent the use of corrupted data, there is no option to correct the readout if the errors that are already present before digitization get detected. Consequently, it is very likely that many profiles cannot be reused after all. For this reason, it is hoped that in the long run, all data published in condensed form as graphs will also be made available to scientists as raw values.

In summary, it was found that digitizing software has become more popular, especially in QSP and PBPK modeling. The presented results indicate that they are a great tool to gather data from graphical representations with excellent accuracy and precision. Moreover, neither user-dependent nor software-dependent relevant confounders could be identified. Although the digitizing time, symbol shape, and software used had a statistically significant influence on digitizing accuracy, the impact on the routine digitizing practice seems negligible. Digitizing a graph more than once did not improve the quality of the readout and thus is redundant. However, it was also found that the greatest danger of incorrectly derived analysis results based on digitized data does not come from the process of digitizing but from pre-existing errors in the published data. Overall, the results of this study

are the results of the first systematic investigation on the accuracy and precision of digitizing software. Hopefully, the derived recommendations as summarized in **Figure 5** may guide and improve the daily digitizing and graph creation practice in the field of QSP and PBPK modeling and eventually enhance the quality of models developed based on digitized readouts.

**Supporting Information.** Supplementary information accompanies this paper on the *CPT: Pharmacometrics & Systems Pharmacology* website ([www.psp-journal.com](http://www.psp-journal.com)).

**Acknowledgments.** The authors thank Ina Schneider and Melanie Titze who carried out the face-to-face interviews and helped design the study and analysis. In addition, the authors thank all the study subjects who diligently digitized our simulated profiles.

**Funding.** This work was funded by the Robert Bosch Stiftung (Stuttgart, Germany), the European Commission Horizon 2020 UPGx grant 668353, a grant from the German Federal Ministry of Education and Research (BMBF 031L0188D), the nanoCELL consortium grant 03XP0196C, the VISION consortium grant 031L0153A and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2180—390900677.

[Correction added on 19 June 2020 after first publication: Funding Information section was included.]

**Conflict of Interest.** The authors declared no competing interests for this work.

**Author Contributions.** J.-G.W., H.B., D.S., M.S., and T.L. wrote the manuscript. J.-G.W., H.B., D.S., M.S., and T.L. designed the research. J.-G.W., H.B., and T.L. performed the research. J.-G.W. and D.S. analyzed the data.

**Data Availability Statement.** The datasets generated and analyzed during the current study as well as the scripts for reproducing the analysis results are available on request from the corresponding author.

1. Lippert, J. *et al.* Open systems pharmacology community—an open access, open source, open science approach to modeling and simulation in pharmaceutical sciences. *CPT Pharmacometrics Syst. Pharmacol.* **8**, 878–882 (2019).
2. Wang, Z. *et al.* Literature mining on pharmacokinetics numerical data: a feasibility study. *J. Biomed. Inform.* **42**, 726–735 (2009).
3. Türk, D. *et al.* Physiologically based pharmacokinetic models for prediction of complex CYP2C8 and OATP1B1 (SLCO1B1) drug-drug-gene interactions: a modeling network of gemfibrozil, repaglinide, pioglitazone, rifampicin, clarithromycin and itraconazole. *Clin. Pharmacokinet.* **58**, 1595–1607 (2019).
4. Britz, H. *et al.* Physiologically-based pharmacokinetic models for CYP1A2 drug-drug interaction prediction: a modeling network of fluvoxamine, theophylline, caffeine, rifampicin, and midazolam. *CPT Pharm. Syst. Pharmacol.* **8**, 296–307 (2019).
5. Hanke, N. *et al.* PBPK models for CYP3A4 and P-gp DDI prediction: a modeling network of rifampicin, itraconazole, clarithromycin, midazolam, alfentanil, and digoxin. *CPT Pharm. Syst. Pharmacol.* **7**, 647–659 (2018).
6. Hanke, N. *et al.* A physiologically based pharmacokinetic (PBPK) parent-metabolite model of the chemotherapeutic zoletarelin doxorubicin-integration of in vitro results, Phase I and Phase II data and model application for drug-drug interaction potential analysis. *Cancer Chemother. Pharmacol.* **81**, 291–304 (2018).
7. Moj, D. *et al.* A comprehensive whole-body physiologically based pharmacokinetic model of dabigatran etexilate, dabigatran and dabigatran glucuronide in healthy adults and renally impaired patients. *Clin. Pharmacokinet.* **58**, 1577–1593 (2019).
8. Schaefer, N., Moj, D., Lehr, T., Schmidt, P.H. & Ramsthaler, F. The feasibility of physiologically based pharmacokinetic modeling in forensic medicine illustrated by the example of morphine. *Int. J. Legal Med.* **132**, 415–424 (2018).

9. Moj, D. *et al.* Clarithromycin, midazolam, and digoxin: application of PBPK modeling to gain new insights into drug-drug interactions and co-medication regimens. *AAPS J.* **19**, 298–312 (2017).
10. GetData Graph Digitizer <<http://getdata-graph-digitizer.com/index.php>>.
11. Connected Researchers Graph digitizer comparison—16 ways to digitize your data <<http://connectedresearchers.com/graph-digitizer-comparison-16-ways-to-digitize-your-data/>>.
12. Engauge Digitizer <<http://markummittchell.github.io/engauge-digitizer/>>.
13. DataThief III <<https://www.datathief.org/>>.
14. Ocampo-Pelland, A.S., Gastonguay, M.R., French, J.F. & Riggs, M.M. Model-based meta-analysis for development of a population-pharmacokinetic (PPK) model for Vitamin D3 and its 25OHD3 metabolite using both individual and arm-level data. *J. Pharmacokinetic. Pharmacodyn.* **43**, 191–206 (2016).
15. Publish or Perish (version 7.12.2505.7183) <<https://harzing.com/resources/publish-or-perish>>.
16. Morley, S.K., Brito, T.V. & Welling, D.T. Measures of model performance based on the log accuracy ratio. *Sp. Weather* **16**, 69–88 (2018).
17. Patro, S.G.K. & Sahu, K.K. Normalization: a preprocessing stage. *IARJSET* 20–22 (2015).
18. Nash, J.C., Zhu, C., Byrd, R., Nosedal, J. & Morales, J.L. `lbfgsb3`: limited memory BFGS minimizer with bounds on parameters <<https://cran.r-project.org/package=lbfgsb3>> (2015).
19. R Development Core Team. R: A Language and Environment for Statistical Computing, Version 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria, 2008) <<https://www.r-project.org>>.
20. RStudio Team. RStudio: Integrated Development Environment for R, Version 1.2.1114 (RStudio, Boston, MA, 2016) <<http://www.rstudio.com/>>.
21. Qt | Cross-platform software development for embedded & desktop <<https://www.qt.io/>>.
22. Winsemius, A. *et al.* Pharmacokinetic interaction between simvastatin and fenofibrate with staggered and simultaneous dosing: Does it matter? *J. Clin. Pharmacol.* **54**, 1038–1047 (2014).
23. Neuvonen, P.J., Kantola, T. & Kivistö, K.T. Simvastatin but not pravastatin is very susceptible to interaction with the CYP3A4 inhibitor itraconazole. *Clin. Pharmacol. Ther.* **63**, 332–341 (1998).
24. Chung, E., Nafziger, A.N., Kazierad, D.J. & Bertino, J.S. Comparison of midazolam and simvastatin as cytochrome P450 3A probes. *Clin. Pharmacol. Ther.* **79**, 350–361 (2006).
25. Knupp, C.A., Brater, D.C., Relue, J. & Barbaiya, R.H. Pharmacokinetics of didanosine and ketoconazole after coadministration to patients seropositive for the human immunodeficiency virus. *J. Clin. Pharmacol.* **33**, 912–927 (1993).

© 2020 The Authors. *CPT: Pharmacometrics & Systems Pharmacology* published by Wiley Periodicals LLC on behalf of American Society for Clinical Pharmacology and Therapeutics. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.