Deciding on Success Criteria for Predictability of Pharmacokinetic Parameters from In Vitro Studies: An Analysis Based on In Vivo Observations

Khaled Abduljalil, Theresa Cain, Helen Humphries, and Amin Rostami-Hodjegan

Simcyp Limited (a Certara company), Sheffield, United Kingdom (K.A., T.C., H.H., A.R-H); and Manchester Pharmacy School, University of Manchester, Manchester, United Kingdom (A.R-H)

Received March 13, 2014; accepted July 2, 2014

ABSTRACT

Prediction accuracy of pharmacokinetic parameters is often assessed using prediction fold error, i.e., being within 2-, 3-, or n-fold of observed values. However, published studies disagree on which fold error represents an accurate prediction. In addition, "observed data" from only one clinical study are often used as the gold standard for in vitro to in vivo extrapolation (IVIVE) studies, despite data being subject to significant interstudy variability and subjective selection from various available reports. The current study involved analysis of published systemic clearance (CL) and volume of distribution at steady state (Vss) values taken from over 200 clinical studies. These parameters were obtained for 17 different drugs after intravenous administration. Data were analyzed with emphasis on the appropriateness to use a parameter value from one particular clinical study to judge the performance of IVIVE and the ability of CL and Vss values obtained from one clinical study to "predict" the same values obtained in a different clinical study using the n-fold criteria for prediction accuracy. The twofold criteria method was of interest because it is widely used in IVIVE predictions. The analysis shows that in some cases the twofold criteria method is an unreasonable expectation when the observed data are obtained from studies with small sample size. A more reasonable approach would allow prediction criteria to include clinical study information such as sample size and the variance of the parameter of interest. A method is proposed that allows the "success" criteria to be linked to the measure of variation in the observed value.

Introduction

In the top-down approach, where the model is derived from clinical data, pharmacometricians use certain criteria for assessing pharmacokinetic predictions or model performance such as goodness-of-fit plots. The prediction of interest is either concentration or response time profiles. Such criteria are not used for bottom-up approach or IVIVE, where the predictions of interest are PK parameters, rather than the profiles. The justification of IVIVE prediction is commonly performed by using the n-fold metric system, and the success of IVIVE and quantitative structure-activity/property relationships prediction is usually assessed by determining the proportion of predictions within a certain fold of previously observed values. However, there is no specification for which fold should be used and thus there is no consistency across different publications. Published articles report their predictions within 1.5-fold (Han et al., 2013), 2-fold (Chen et al., 2012), 3-fold (Gibson et al., 2009), and 5-fold (Gombar and Hall, 2013) of the observed value. It is worth mentioning here that the term "observed values" in the bottom-up approach should be distinguished from that in the top-down approach. In the bottom-up approach, the term "observed values" refers to the collected PK parameter values (e.g., systemic clearance [CL] or volume of distribution at steady state [Vss]), whereas in the top-down approach the observed values usually refer to the measured concentration in plasma. The aim of the top-down approach is not to predict the PK/pharmacodynamics parameter itself but to predict the impact of the PK parameter on concentration, amount, and or response in any biologic matrix such as plasma, urine, blood, or tissue.

The observed data in the case of IVIVE predictions are the PK parameters; the values that are considered to be gold standard data. However, these values include uncertainties due to bias or imprecision in calculating them from incomplete data and also because of the inherent variability. Reported clinical PK parameters are subject to both inter- and intrastudy variability, which stems from different sources such as ethnicity; genetic variation in cytochrome P450 metabolizing enzymes (Tucker, 1994) such as CYP2D6 (Dorne et al., 2002; Abduljalil et al., 2010; Thompson et al., 2011), CYP2C9 (Loebstein et al., 2001; Borgiani et al., 2007), and CYP2C19 (Dorne et al., 2003; Crettol et al., 2005); comedication; health status; different assay and analysis; environmental factors; and clinical settings of the study such as sample size.

Intersubstudy variability in the clinical PK parameters may bias the assessment of IVIVE prediction accuracy when predicted PK parameters are compared with data from only one particular clinical study. It is also desirable if the predicted variability matches that of the observed in the assessment of the predictability. These multiple sources of variability create a challenge for comparing PK parameters between clinical studies themselves, and this challenge increases significantly when one PK value is chosen to assess the performance of IVIVE.

ABBR EV IATIONS: CIs, confidence intervals; CL, systemic clearance (after intravenous administration); CV, coefficient of variation; IVIVE, in vitro to in vivo extrapolation; PK, pharmacokinetics; Vss, volume of distribution at steady state (after intravenous administration).
A significant amount of attention has been paid to the prediction of PK parameters and their dispersion from in vitro data (Howgate et al., 2006; Inoue et al., 2006; Jamei et al., 2009; Cubitt et al., 2011) the accuracy of using PK parameters obtained from one clinical study to "predict" the same parameter obtained in a different clinical study using the n-fold metric system has not been investigated.

The aims of this paper are 1) to investigate whether parameters from one particular clinical study can be used to judge the performance of IVIVE and if this is determined to be appropriate, how to assess which clinical study the PK parameters should be taken from; 2) to investigate whether PK parameters obtained from one clinical study can "predict" the same values obtained in a different clinical study within a certain fold; 3) to propose an improved method for defining the prediction criteria that considers sample size and the variance around the PK parameter of interest.

**Materials and Methods**

**Clinical PK Parameter Data Collection**

**Data Sources.** Structured literature searches were carried out using Medline for the parameters CL and Vss of 17 compounds (Table 1) strictly after intravenous doses. The Vss data after intravenous administration were collected for fewer drugs because the main focus was on the ability of the n-fold metric system to predict parameters rather than the actual predictability of the parameters. There were no criteria about compound selection other than data availability. No language or date restriction was applied, but article titles and abstracts were screened to maintain the focus of the search upon these two parameters. Manual search of references lists from selected articles complemented the data collection process. Data were extracted and entered into an Excel spreadsheet, which was subsequently checked before analysis.

Data inclusion criteria were adult healthy Caucasian individuals. No restriction on sex or maximum age was applied to the inclusion criteria because this is not applied during IVIVE. The exclusion criteria were underlying health conditions known to affect the pharmacokinetic parameters. For example, studies were excluded if they reported health conditions such as renal insufficiency, smoking, cirrhosis, pregnancy, and obesity. Studies that reported only central tendency values without variability were also excluded from this study.

**Clinical PK Parameter Data Analysis.**

Data analysis was performed using Microsoft Excel 2010. CL and Vss values were reported with various units and obtained via compartmental and noncompartmental PK analysis. To enable comparison between studies, all CL and Vss units were converted to liters per hour and liters per kilogram, respectively. A reference value of 70 kg body weight was assumed if the mean weight in the original paper was not reported.

In this analysis, all reported CL values from different studies were put together for each compound. A CL value among these values was randomly selected and assumed to be the "true" value. This value was then compared with the remaining values that were considered to be "predictions" of the "true" value. The procedure was performed for each CL value within the collected values for the drug X to allow each value equal chance to represent the true value. This strategy of analysis was done for both CL and Vss separately and for each of the 17 drugs in turn. "Predictions" were plotted against the "true" values for each parameter for each compound, and the percentage of "predictions" failing to be within 1.25-, 1.5-, 2.0-, 2.5-, and 3-fold limits of the "true" value were calculated.

**Hypothetical Data Simulation to Develop an Alternative Success Criteria.**

Trial mean clearance values were simulated using the statistical software package R version 2.12 (www.r-project.org) to assess the suitability of the twofold limits for varying sample sizes and population CV% values. A population geometric mean CL value of 100 l/h was assumed. This value was selected for convenience, because comparing trials with the same actual mean demonstrates the ability of the twofold prediction criteria to accurately predict the means. Trials were generated using this geometric mean, assuming sample sizes of between 5 and 20 and CVs between 10 and 150%. For each combination of sample size and CV%, 100,000 trials were simulated, and the mean of each trial was then calculated.

The CL was assumed to have a lognormal distribution, \( CL = \ln(100) \sigma \), where \( \sigma \) is the standard deviation of the data on the natural log scale (scale parameter of the lognormal distribution) and is calculated from the CV% value using the equation

\[
\sigma = \sqrt{\ln \left( \frac{\text{CV}\%}{100} \right)^2 + 1}
\]

The trial mean clearance values were simulated in R using the predefined function using for generating values from a lognormal distribution. The percentage of trial means outside of the twofold limits of the population geometric mean were then calculated for each combination of sample size and CV% value.

The trial means generated using a sample size of 10 were then investigated further by plotting them on a graph and visually assessing their distribution relative to both the twofold limits, the 95% confidence intervals around the geometric mean CL, and the new proposed metric system, which is the 99.998%
confidence interval around the geometric mean CL. The 95% geometric confidence intervals (CIs) are calculated using eq. 2:

\[
\exp \left[ \ln(\bar{x}) \pm 1.96 \frac{\sigma}{\sqrt{N}} \right]
\]

where \(\ln(\bar{x})\) is the natural logarithm of the mean value. The 99.998% geometric confidence intervals were proposed as a new metric system to represent the majority of the population distribution of the sample mean without having CIs between 0 and infinity. The 99.998% geometric CIs are calculated using eq. 3:

\[
\exp \left[ \ln(\bar{x}) \pm 4.26 \frac{\sigma}{\sqrt{N}} \right]
\]

The scale parameter \(\sigma\) value was calculated from the assumed CV\% and sample size \(N\).

These new metrics using the 99.998% confidence intervals can be considered in terms of fold limits instead of the standard twofold limits. The twofold limits can be written as 0.5\( \leq x \leq 2\) for geometric mean \(\bar{x}\). In general, the fold limits can be written as \(Bx \leq X \leq Ax\), where the predicted mean is accepted if it is within the upper (\(Ax\)) and lower (\(Bx\)) fold limits of the “true” geometric mean (\(\bar{x}\)). These general upper and lower fold limits can be set equal to the upper and lower 99.998% geometric mean CIs as shown in eq. 4 and eq. 5, respectively,

\[
\bar{x}A = \exp \left[ \ln(\bar{x}) + 4.26 \frac{\sigma}{\sqrt{N}} \right]
\]

and

\[
\bar{x}B = \exp \left[ \ln(\bar{x}) - 4.26 \frac{\sigma}{\sqrt{N}} \right]
\]

where \(\sigma\) is the scale calculated (using eq. 1) from the reported CV\% and \(N\) is the sample size of the reported mean in the clinical study. Equations 4 and 5 are considered as a new fold metric approach to assess the ability of a parameter in predicting new parameters.

Finally, this new proposed approach was used to evaluate the prediction of CL and V\(ss\) in comparison with the twofold metric assessment as the cut-off. A prediction was considered successful and acceptable if the predicted value was within the limits of the assessment method of interest. The percentages of data within the twofold or 99.998% CIs were calculated by counting how many predictions were inside these higher and lower boundaries as a percentage of the whole data set.

### Results

**Clinical PK Parameter Data Analysis**

The data collected include wide spectral compounds that undergo extensive metabolism, such as midazolam, or are eliminated mainly by renal excretion, such as furosemide. The collected data are presented in Table 1 (see Supplemental Table A for the full list of the collected references).

When comparing CL values from each study with the CL values in all other studies using the twofold metric for a given compound, it was observed that 13 of the 17 drugs had CL values obtained from at least one clinical study that was outside the twofold limits of a CL value from another study. The percentage of mean CL values outside the twofold limits varied between 2\% (lidocaine, theophylline, and antipyrine) and 18\% (digoxin). Only 5 of the 17 drugs had a percentage error of \(\geq 10\%\) (propofol, midazolam, digoxin, alfentanil, and lorazepam). This percentage error was reduced to \(<5\%\) for all compounds when the threefold accuracy criteria was considered (Table 2).

\(V_{ss}\) values after intravenous administration were available for 11 of the 17 drugs. Of the 11 compounds, 5 had mean \(V_{ss}\) values that were outside the twofold limits of a value obtained from another study. The range in percentage error for "predictions" was between 2\% (lidocaine and diazepam) and 22\% (midazolam). Only 2 of the 11 drugs had a percentage error of \(\geq 10\%\) (midazolam and digoxin). This error was reduced to \(<5\%\) for all compounds when the threefold accuracy criteria was considered (Table 2).

Plots of each “true” value for CL and \(V_{ss}\) against all other "predicted" values are presented in Figs. 1 and 2, respectively, with the twofold system prediction limits. These compounds were randomly selected to represent compounds of high, medium, and low variability.

**Hypothetical Data Simulation to Develop an Alternative Success Criteria**

For each scale parameter value of the lognormal distribution and sample size, 100,000 trials were generated for an assumed CL population geometric mean of 100 l/h. The percentage of trial means outside the 99.998% CIs of the population geometric mean is presented in Supplemental Table B by sample size and CV\%. This is also presented graphically in Fig. 3. The percentage of predicted means outside of the 99.998% CIs increases as the sample size decreases and as the CV\% of the PK parameter increases.

**TABLE 2**

Percentages of predictions outside the corresponding \(n\)-fold

<table>
<thead>
<tr>
<th>Compound</th>
<th>CL</th>
<th>V(ss)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.25-fold</td>
<td>1.5-fold</td>
</tr>
<tr>
<td>Tolnafoxin</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>Propranolol</td>
<td>53</td>
<td>29</td>
</tr>
<tr>
<td>Cisatracurium</td>
<td>36</td>
<td>8</td>
</tr>
<tr>
<td>Acetaminophen</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>48</td>
<td>20</td>
</tr>
<tr>
<td>Propofol</td>
<td>56</td>
<td>35</td>
</tr>
<tr>
<td>Metronidazole</td>
<td>50</td>
<td>27</td>
</tr>
<tr>
<td>Theophylline</td>
<td>44</td>
<td>19</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>43</td>
<td>17</td>
</tr>
<tr>
<td>Midazolam</td>
<td>56</td>
<td>32</td>
</tr>
<tr>
<td>Antipyrine</td>
<td>54</td>
<td>26</td>
</tr>
<tr>
<td>Digoxin</td>
<td>63</td>
<td>37</td>
</tr>
<tr>
<td>Furosemide</td>
<td>55</td>
<td>28</td>
</tr>
<tr>
<td>Lorazepam</td>
<td>59</td>
<td>36</td>
</tr>
<tr>
<td>Diazepam</td>
<td>56</td>
<td>30</td>
</tr>
<tr>
<td>Thiopental</td>
<td>34</td>
<td>7</td>
</tr>
<tr>
<td>Alfentanil</td>
<td>68</td>
<td>45</td>
</tr>
</tbody>
</table>

NA, not applicable (data were not available).
Figure 4 presents a plot of the simulated trial means using a sample size of 10 plotted against the CV%. In this figure, three criteria, the twofold limits, 95% CIs, and the new proposed metric system using the 99.998% geometric CIs, are plotted for comparison. As the scale parameter (and therefore CV%) increases, the variability of the sample trial means increases as expected. For smaller scale parameter values (and therefore smaller CV%), the twofold limits appear to be too wide to be used as prediction intervals and potentially allow values to be accepted as good predictions, which should not be, i.e., false positives. In contrast, for the larger scale parameter values (and therefore larger CV%), not all trial means are within the twofold limits and therefore values that are good predictions will not be accepted, i.e., false negatives. The proposed method using the 99.998% CIs appear to include most sample trial means and reduce the chance of either accepting clinically irrelevant values for the smaller values of CV% or rejecting clinically relevant values for the larger values of CV%.

Comparison of Twofold versus Alternative Success Criteria

A comparison between the twofold and alternative success criteria is presented in Table 3 and Fig. 5. The new proposed method using the 99.998% confidence intervals results in an increased percentage of "predicted" CL or Vss values outside of the proposed limits in comparison with the twofold criteria, because it is limited by the sample size and variability of the compound. With acetaminophen and propranolol as examples, it can be seen that the boundaries of the new proposed method are contained within the twofold acceptance boundaries (Fig. 5).

Discussion

This study investigated the impact of in vivo variability of PK parameters and the metric system commonly used to assess prediction accuracy for compounds with a wide range of linear disposition properties.
There is evidence to suggest that in a clinical setting, the twofold prediction error in drug PK parameters is acceptable for most drugs; however, acceptance criteria will vary between drugs. There is high variability in the PK parameters for some compounds, whereas a few have relatively low variability (Table 1). For drugs with high variability, such as digoxin and midazolam, the 2.5-fold metric system may be acceptable. For drugs with intermediate variability like diazepam, the twofold system seems appropriate. For drugs with low variable PK parameters, such as talinolol and acetaminophen, the twofold criteria boundaries are wide and a tighter 1.5-fold could be appropriate.

In principle the prediction criteria described in the manuscript for clearance is also applicable for other PK parameters such as volume of distribution. The predictability of volume of distribution between studies in this analysis was better than that of clearance, because there was less variability in the case of distribution volume than in clearance. This is due to the fact that clearance depends on area under the curve, whereas Vss calculation depends mainly on the first few samples.

A major limitation of the twofold criteria is that it handles data that come from different studies equally, irrespective of the sample size. For example, it can accept values from studies with small sample sizes as good predictions if they are within twofold boundary while rejecting the predictions of values from studies with high sample sizes if they are outside the twofold boundary (see digoxin example in Fig. 1). Previously, this system received criticism in the field of drug interactions because it results in a potential bias toward successful prediction at lower interaction levels and can bias any assessment of different drug-drug interaction prediction algorithms if databases contain a large proportion of interactions in the lower range of interaction (Guest et al., 2011).

Simulation results presented in Fig. 3 show that for a sample size of 10, the percentage of means (of 100,000 trials) outside the 99.998% CIs limits increases as the CV% value increases. This is particularly the case as the sample size decreases, as shown in Fig. 3, which presents the case for sample sizes between 5 and 20. The greatest percentage of trial means outside the limits is 41% for a sample size of 5 and a CV% of 150 (see Supplemental Table B). However, for CV% of 10 and 20, all simulated trial means are within the limits. This suggests that the percentage within the limits depends on both the sample size and the CV% of the population. These findings show that the limits should be related to the CV% of the population and the size of the sample used in the prediction.
Looking at the simulated dataset for a sample size of 10 (Fig. 4), the twofold limits for smaller values of CV% appear to be too wide to be used as prediction intervals and could allow the acceptance of a value that is not a true representation of that population, a high false positive rate. Likewise, the number of simulated trial means outside the twofold limit for larger CV% values suggests these limits are too small and could allow values not to be accepted as representative of the population when they are true values, a high false negative rate. Both sets of geometric CIs appear to be a similar shape to the distribution of simulated means, suggesting the acceptance limits should be related to these CIs (Fig. 3).

The new proposed method using the 99.998% CIs appear to include most of the simulated trial means, and if these limits were included instead of the twofold limits it would reduce both the false negative and false positive rates of prediction accuracy. If the 99.998% CIs were used as the fold limits, it could be shown that the fold limits depend on both the sample size and CV% as shown in Fig. 4.

To apply the new method using 99.998% CIs, one needs to calculate the sigma based on CV and sample size from the reference in vivo study using eq. 1 and substitute for sigma in eqs. 4 and 5, which calculate the sigma based on CV and sample size from the reference in vivo study. It should be pointed out here that this work was used as the fold limits, it could be shown that the fold limits depend on both the sample size and CV% as shown in Fig. 4.

The comparison between the twofold and new metric systems acceptance criteria was given in Table 3 for all compounds used in the analysis and in Fig. 5 for two selected compounds. Taking the CL of propranolol as an example, the collected data gave an overall CL of 21.6 l/h (46%) mean CL value of 152 nor 15 l/h has been reported. They are unlikely to be clinically relevant values for propranolol systemic CL in healthy individuals either.

On the other hand, the new proposed method using 99.99% geometric confidence intervals accepts a prediction range of 60.0–96.1 l/h for the observed CL value of 76 l/h with the S.D. of 14.7 l/h that came from a study with a sample size of 12 individuals (Cheymol et al., 1997). Likewise, it accepts a prediction range of 23.34–38.43 l/h for the observed CL value of 30 l/h with the S.D. of 5 l/h that came from a study with a sample size of 8 individuals (Castleden and George, 1979). According to eqs. 4 and 5, the prediction of the new proposed method will change if either the sample size or the dispersion parameters are changed.

Limitations of this analysis are that the analysis was carried out for selected studies based on some selection criteria (see Materials and Methods) and did not consider further subgrouping studies according to the assay method or some demographics like sex, body weight, etc., because such covariates are not available in all publications. The collected data for those compounds may not cover the actual variability of compounds under study. It should be pointed out here that this work

<table>
<thead>
<tr>
<th>Compound</th>
<th>CL 2-Fold</th>
<th>Proposed System</th>
<th>Vss 2-Fold</th>
<th>Proposed System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talinolol</td>
<td>0</td>
<td>27</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Propranolol</td>
<td>6</td>
<td>54</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Cisatracurium</td>
<td>0</td>
<td>44</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Acetaminophen</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>2</td>
<td>39</td>
<td>2</td>
<td>42</td>
</tr>
<tr>
<td>Propofol</td>
<td>13</td>
<td>58</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Metronidazole</td>
<td>5</td>
<td>47</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>Theophylline</td>
<td>2</td>
<td>26</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>3</td>
<td>38</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>Midazolam</td>
<td>11</td>
<td>54</td>
<td>22</td>
<td>59</td>
</tr>
<tr>
<td>Antipyrine</td>
<td>2</td>
<td>46</td>
<td>0</td>
<td>41</td>
</tr>
<tr>
<td>Digeoxin</td>
<td>18</td>
<td>46</td>
<td>13</td>
<td>36</td>
</tr>
<tr>
<td>Furosemide</td>
<td>6</td>
<td>29</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>Thiopental</td>
<td>0</td>
<td>15</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>Lorazepam</td>
<td>10</td>
<td>37</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>Diazepam</td>
<td>5</td>
<td>32</td>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>Atalantil</td>
<td>16</td>
<td>38</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

NA, not applicable (data were not available).

Fig. 5. Comparison between the twofold versus the new prediction accuracy criteria for propranolol (top plot) and acetaminophen (bottom plot) CL. Red dots are the range in "true" CL values reported by the different clinical studies. The identity line represents the same CL values plotted in numerical order (the "predicted" CL values). Upper and lower twofold acceptance boundaries are given as solid lines. The new method boundaries are given as dashes (-).
was done in the limited case of known linear drugs, and the situation might be in need of further assessment in the case of nonlinear kinetics.

In conclusion, the arbitrary twofold system shows a wide range of acceptance for low variable drugs and vice versa. The discussed alternative prediction accuracy criteria will allow the prediction to take into account clinical settings and parameter variability.

Acknowledgments

The authors thank James Kay for assistance in gathering the data and Eleanor Savill for preparing the manuscript submission. The authors also thank the students in the Modelling & Simulation course at the University of Sheffield and the University of Manchester for their initial contributions to the analysis.

Authorship Contributions

Participated in research design: Abduljalil, Humphries, Rostami-Hodjegan.

Conducted experiments: Abduljalil, Humphries.

Performed data analysis: Abduljalil, Cain.

Wrote or contributed to the writing of the manuscript: Abduljalil, Cain, Humphries, Rostami-Hodjegan.

References


Address correspondence to: Khaled Abduljalil, Simcyp Limited, Blades Enterprise Centre, John Street, Sheffield, S2 4SU, UK. E-mail: Khaled.Abduljalil@certara.com