Received: 2 December 2021 | Revised: 14 March 2022 | Accepted: 16 March 2022

DOI: 10.1002/psp4.12790

TUTORIAL

A quantitative approach to the choice of number of samples for percentile estimation in bootstrap and visual predictive check analyses

E. Niclas Jonsson | **Joakim Nyberg**

Pharmetheus AB, Uppsala, Sweden

Correspondence
E. Niclas Jonsson, Pharmetheus AB, Dag Hammarskjölds väg 36B, SE-752 37 Uppsala, Sweden.
Email: niclas.jonsson@pharmetheus.com

Funding information
No funding was received for this work.

Abstract

Understanding the uncertainty in parameter estimates or in derived secondary variables is important in all data analysis activities. In pharmacometrics, this is often done based on the standard errors from the variance–covariance matrix of the estimates. Confidence intervals derived in this way are by definition symmetrical, which may lead to implausible outcomes, and will require translation to generate uncertainties in derived variables. An often-used alternative is numerical percentile estimation by, for example, nonparametric bootstraps to circumvent these issues. Visual predictive checks (VPCs), which is a commonly used model diagnostic tool in pharmacometric analyses, also rely on the estimation of percentiles through numerical approaches. Given the cost in terms of run times and processing times for these methods, it is important to consider the trade-off between the number of bootstrap samples or simulated data sets in the VPCs, to the increase in precision related to a large number of bootstrap samples or simulated data sets. The objective with this tutorial is to provide a quantitative framework for assessing the precision in estimated percentile limits in bootstrap and visual predictive checks analyses to facilitate an informed choice of confidence interval width, number of bootstrap samples/simulated data sets, and required level of precision.

INTRODUCTION

Understanding the uncertainty in parameter estimates or in derived secondary variables is important in all data analysis activities. In pharmacometrics, this is often done based on the standard errors from the variance–covariance matrix of the estimates. Confidence intervals derived this way are per definition depending on the assumed distribution and often symmetrical, which may lead to implausible outcomes (for example, that the confidence interval include negative elimination rates), and will require translation to generate uncertainties in derived variables. An often-used alternative is numerical percentile estimation by, for example, nonparametric bootstraps[1] to circumvent the issues with the standard errors from the variance–covariance matrix. Visual predictive checks[2] (VPCs), which is a commonly used model diagnostic tool in pharmacometric analyses, also rely on the estimation of percentiles through numerical approaches.

When the nonparametric bootstrap is used to estimate confidence interval limits, the procedure involves re-estimation of the model on (many) resampled data sets to create a multivariate distribution of parameter estimates from which the marginal percentiles of interest can be derived. In VPCs, percentile-based confidence intervals are derived through (many) simulations from the model, which are then compared to the corresponding percentiles from the real data set.

The bootstrap confidence intervals are typically used for inferential purposes, for example, to report the precision in parameter estimates or covariate effects in Forest plots.[3] VPCs are used as diagnostics during model development or as a way to qualify a final model. In both cases, the reliability of the conclusions depends on the precision with which the percentile limits have been determined, and the precision depends on the number of samples, bootstrap or simulated, that is used and the percentile that is estimated. The more samples used (bootstrap resamples or simulations), the higher precision and more certain conclusions, although too few samples lead to more uncertain conclusions. In addition, the more extreme percentiles that are estimated (for example, the 2.5th rather than the 5th percentile) the more samples are needed for the same level of precision. It is of course tempting to use a very high number of samples to avoid the small sample uncertainty but because there is a computational cost associated with each sample this strategy may not be feasible. In particular, the nonparametric bootstrap is computationally expensive because each bootstrap sample involves a re-estimation of the model, but run-times and/or disc space requirements may also be prohibitive for VPCs even if they are based solely on simulations. A practical consideration to keep in mind is that bootstraps, and to some extent also VPCs, are often end-of-analysis activities meaning that if there are time constraints, which is often the case in drug development, time may not allow a large number of samples and may make these methods impractical or be forced to be performed with too low precision. However, bootstraps and VPCs may also used as diagnostics during the model development, and given the computational burden involved it may be tempting to use fewer samples to save time and thereby making model development decisions that are not well founded. The challenge in either case is to find the balance between a sufficient degree of precision in the estimated percentiles so that decisions and inference can be made with adequate certainty, and the practical constraints imposed by lengthy computations.

In the pharmacometric field, the required number of nonparametric bootstrap samples appear to have converged at $N = 1000$. This statement is based on some recent publications[4–7] and on an empirical investigation on the stability of the bootstrap estimates.[8] In these investigations, the nonparametric bootstrap was used to estimate 95% confidence intervals (i.e., the 2.5th and 97.5th percentiles). There is no corresponding systematic investigation of the impact of the number of samples for VPC analyses but in the original publication on the prediction corrected VPC by Bergstrand et al.[2] 1000 VPC samples were used.

In practice, the choice of the number of samples used for percentile estimation is likely to be a mix between practical constraints, a perceived need to be on the "safe side" and a tendency to do what others have done. This means that across presentations of results based on percentile estimation, there will be a mix of more imprecise percentile limits (when a small sample number was used to manage practical constraints) and very precise but overly inefficient analyses (when a large sample number was used). Such differences will obviously occur between analyses but is likely to also occur within analyses (as will be illustrated below), and it is hard for readers of analysis reports and papers to understand and assess on what level of precision conclusions are drawn and/or models are developed.

The objective with this tutorial is to provide a quantitative framework for assessing the precision in estimated percentile limits in bootstrap and VPC analyses, to facilitate an informed choice between confidence interval width, the number of samples, and the required level of precision.

In the following, the focus will first be on the nonparametric bootstrap, with which the quantitative framework for the precision in estimated percentile limits will be explained and exemplified. There will also be a section on an alternative use of the bootstrap to derive uncertainties which does not require as many bootstrap samples. This will be followed by a section of the VPC and how the framework established in the bootstrap section can be used also in this setting. The tutorial will be concluded with a section with some concrete recommendations as well concluding remarks.

THE NONPARAMETRIC BOOTSTRAP

The bootstrap was originally suggested by Efron[1] as a way to assess the uncertainty of parameters without making strong distributional assumptions. There are many ways the general principles of the bootstrap can be applied to pharmacometric models but the focus of this work is on the nonparametric case bootstrap (i.e., random sampling of complete individuals),[6] with or without stratification.

In the nonparametric case bootstrap, complete individuals (all data records from an individual) are sampled

from the analysis data set with replacement. Typically, the number of sampled individuals is the same as the original number of individuals. The sampling is repeated N times—this is the number of bootstrap samples discussed in the Introduction. The model is fit to each of the N sampled data sets and the N sets of parameter estimates make up the multivariate distribution from which the bootstrap results is derived. Figure 1 shows a histogram of 1000 bootstrap parameter estimates. The percentile estimates from a bootstrap analysis are derived from distributions like this. The x-axis indicates some percentiles, and, for example, the limits of a 90% confidence interval are defined by the 5th and 95th percentiles, as indicated in Figure 1. In addition to the confidence interval for the model parameter, Figure 1 also indicates the (schematic) confidence interval for the percentile limits (p5% and p90%), in other words the confidence interval around the limits of another confidence interval.

To separate the two confidence intervals, the acronym CI will be used for "higher" level confidence interval, for example, the confidence interval for the estimate of clearance, whereas the confidence interval for the estimates of the upper and lower percentile limits will be denoted $CI_{p,lo}$ and $CI_{p,up}$. If the percentile limit is irrelevant for the reasoning the lo and up may be dropped so that CI_p refers to either of $CI_{p,lo}$ and $CI_{p,up}$.

We will also denote the lower and upper limits of CI with CI_{lo} and CI_{up}, and the upper and lower limits of $CI_{p,lo}$ and $CI_{p,up}$ with $CI_{p,lo,lo}$ and $CI_{p,lo,up}$, and $CI_{p,up,lo}$, and $CI_{p,up,up}$, respectively.

The width of CI is CI width $= CI_{up} - CI_{lo}$, for example, $CI_{up} - CI_{lo} = 0.975–0.025 = 0.95$ for a 95% confidence interval.

Finally, we will define the width of $CI_{p,lo}$ and $CI_{p,up}$ as $CI_{p,lo}$ width $= CI_{p,lo,up} - CI_{p,lo,lo}$ and $CI_{p,up}$ width $= CI_{p,up,up} - CI_{p,up,lo}$, respectively, similarly to the CI width above. The notation is visualized in Figure 2.

The width of the CI_p intervals define the precision with which the percentile limits are estimated and can be calculated using the standard error (SE) for a binomial fraction p (p = percentile/100) and the number of bootstrap samples. Under normality assumptions and provided the p of interest is not too close to the extremes (i.e., 0 or 1), the SEs can be computed according to Equation 1 and the corresponding relative SE (RSE) according to Equation 2.[9]

$$SE_p = \sqrt{\frac{p(1-p)}{N}} \tag{1}$$

$$RSE_p = \frac{\sqrt{\frac{p(1-p)}{N}}}{p} \tag{2}$$

p is the percentile/100, N the number of bootstrap samples and SE_p and RSE_p are the SE and RSE of p, respectively.

The corresponding CI_p is given by Equation 3.

$$100 \times \left(p - z_{\alpha/2}SE_p,\ p + z_{\alpha/2}SE_p\right) \tag{3}$$

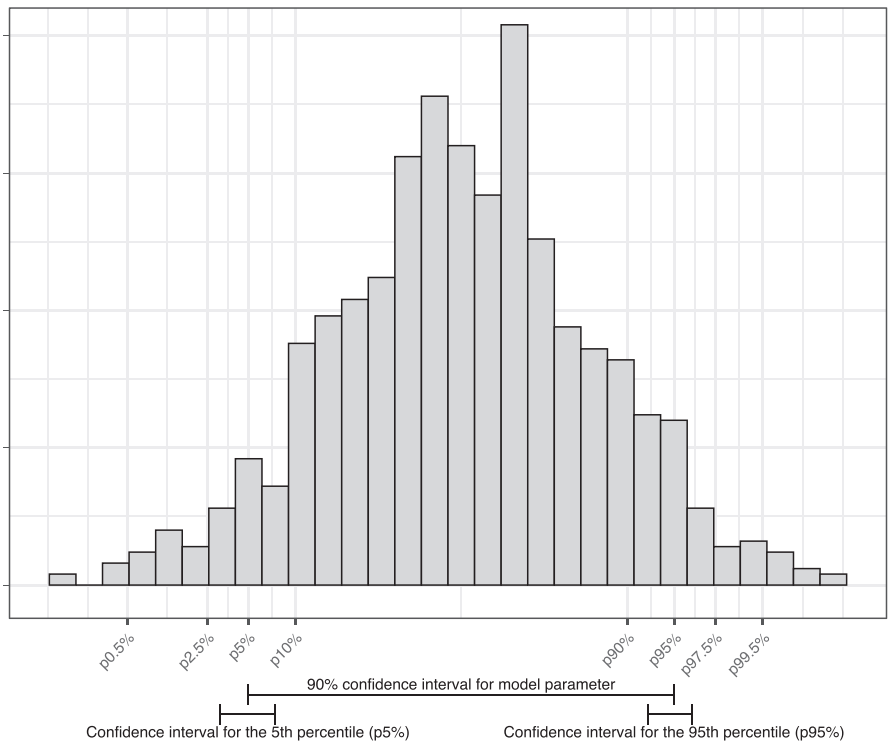

FIGURE 1 Histogram of 1000 bootstrap estimates of a model parameter (e.g., clearance). The scale on the x-axis indicates the percentiles of the distribution (p0.5–p99.5%). Indicated is also the 90% confidence interval for the model parameter, based on the 5th and 95th percentile of the bootstrap distribution, as well as the confidence intervals around the percentile estimates

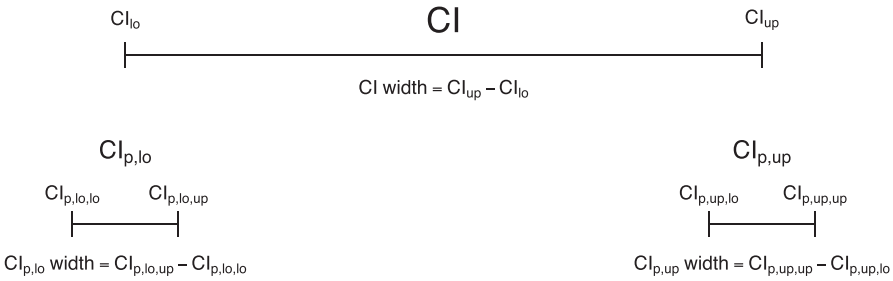

FIGURE 2 Visualization of the notation used to refer to the various confidence intervals (CI) and confidence interval limits

α is the desired width of the CI_p and $z_{\alpha/2}$ is the corresponding z values from the standard normal distribution (for example, 1.96 for a 95% width of CI_p).

The normality assumption involved in the above formulas are not uncontroversial, especially as p approaches 0 or 1, and many alternative ways to compute confidence intervals for binomial fractions have been proposed. Brown et al.[10] evaluates a number of these alternatives and concludes that the correction suggested by Agresti and Coull[11] behaves as well as the other investigated alternatives for N greater than 40 while being fairly straightforward to present.

In brief, the Agresti-Coull correction involves replacing p and N in Equation 1 with other, adjusted, values that produce confidence intervals with better coverage probabilities.

Let x be the number of "successes" (a "success" in our case would be the number of bootstrap samples that results in a value outside the defined percentile interval) and let N be the number of "tries" (the number of bootstrap samples), then the associated probability p is given by Equation 4.

$$p = x/N \tag{4}$$

The corrected x (=x*) and N (=N*) are given by Equation 5 and the corrected p (=p*) by Equation 6.

$$x^* = x + \frac{z_{\alpha/2}^2}{2} \quad \text{and} \quad N^* = N + z_{\alpha/2}^2 \tag{5}$$

α is the desired width of the confidence interval.

$$p^* = x^*/N^* \tag{6}$$

The corrected versions of our SE_p, RSE_p, and CI_p are given by replacing p and N with p* and N* in Equations 1–3.

In the following, all calculations and results will use the Agresti-Coull adjusted p and N but for simplicity the * will be dropped from the notation.

The relationship among N, CI width, and precision in CI_p

From Equation 1, it is clear that for a given true value of p, the smaller N is, the larger SE_p (and therefore CI_p) becomes, or in other words, the more uncertain the estimate of p is. This is illustrated in Figure 3 where results from 3 × 3 bootstrap experiments are shown. All experiments are based on the same data set of 100 random numbers and the objective is to compute the 90% confidence interval around the mean. In the first experiment, this is done by sampling 100 values with replacement from the original data set 20 times and compute the mean from each of these 20 data sets. The 20 means from the first experiment is displayed in the top left corner of Figure 3. The second and third experiments are identical to the first except that 100 and 1000 data sets were sampled from the original data set, respectively. The results are shown in the top row in Figure 3. The three experiments were repeated twice (middle and bottom rows in Figure 3). Each of the panels in Figure 3 shows the estimated 90% confidence intervals as dashed blue lines. Because the confidence interval of interest should cover 90%, the target percentiles are (1 – 0.9)/2 = 5% and 1 − (1 – 0.9)/2 = 95%, indicated on the x-axes as p5% and p95%, respectively.

In the experiments with N = 20, the histograms of the computed means vary substantially between experiments as do the estimated 90% confidence interval limits. When N = 1000, on the other hand, the histograms and the estimated 90% confidence interval limits display considerable similarity. The N = 100 experiments fall in between the N = 20 and N = 1000 experiments, but, at least in these three experiments, appears to provide a lot less variable results compared to N = 20. The N = 1000 experiments are less variable than the N = 100 but, perhaps, not to the extent one would expect given the factor 10 increase in N. The reduced variability going from N = 20 to N = 100 (a factor 5) is much more pronounced.

This is more systematically illustrated in Figure 4. In the left panel, the RSE in the CI_p interval limit for different CI widths are plotted versus the number of bootstrap samples. The RSEs initially decrease rapidly as N increases but at higher Ns the rate of decrease in RSE slows down. There is also a higher cost in terms of N to decrease the RSE in the wider 99% CI compared to the narrower CIs. The right panel illustrates the impact of the size of the CI_p (via the Agresti-Coull correction). In the left, the results are conditioned on a desired CI_p of 90%. In the right panel,

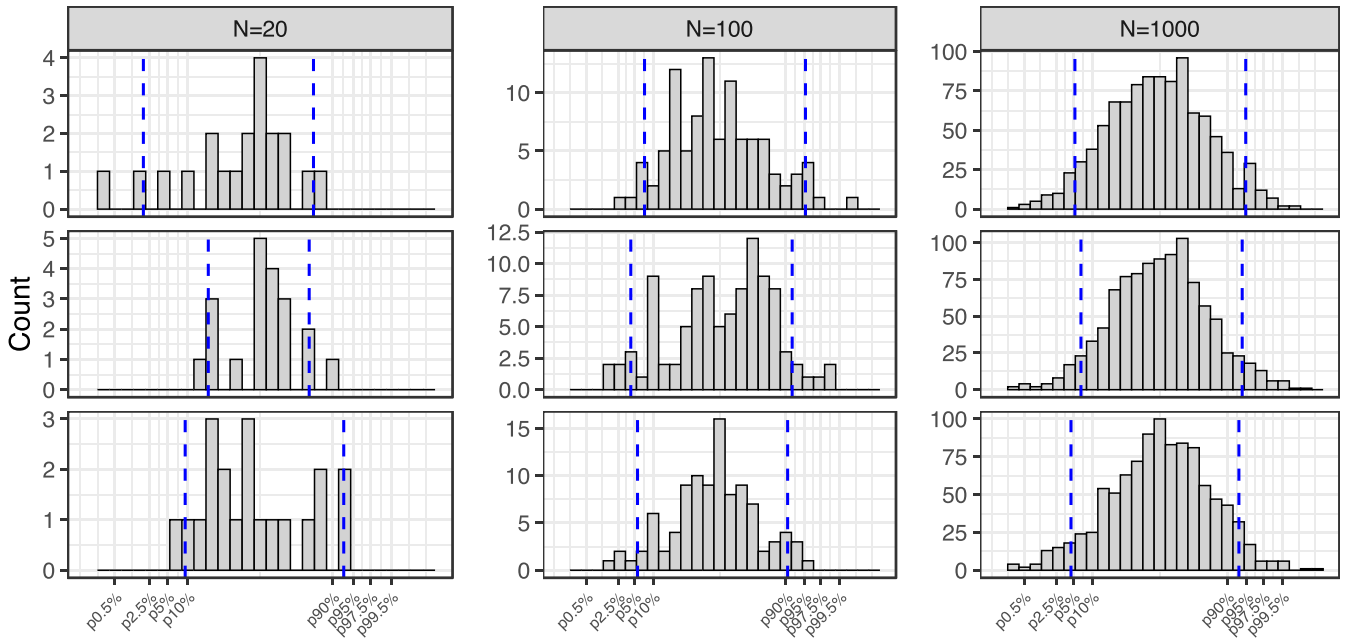

FIGURE 3 Histograms of means from bootstrap samples. The columns strips indicate the sample size and the rows are different replicates. The dashed blue lines indicate the 5th and 95th percentile. See the text for details

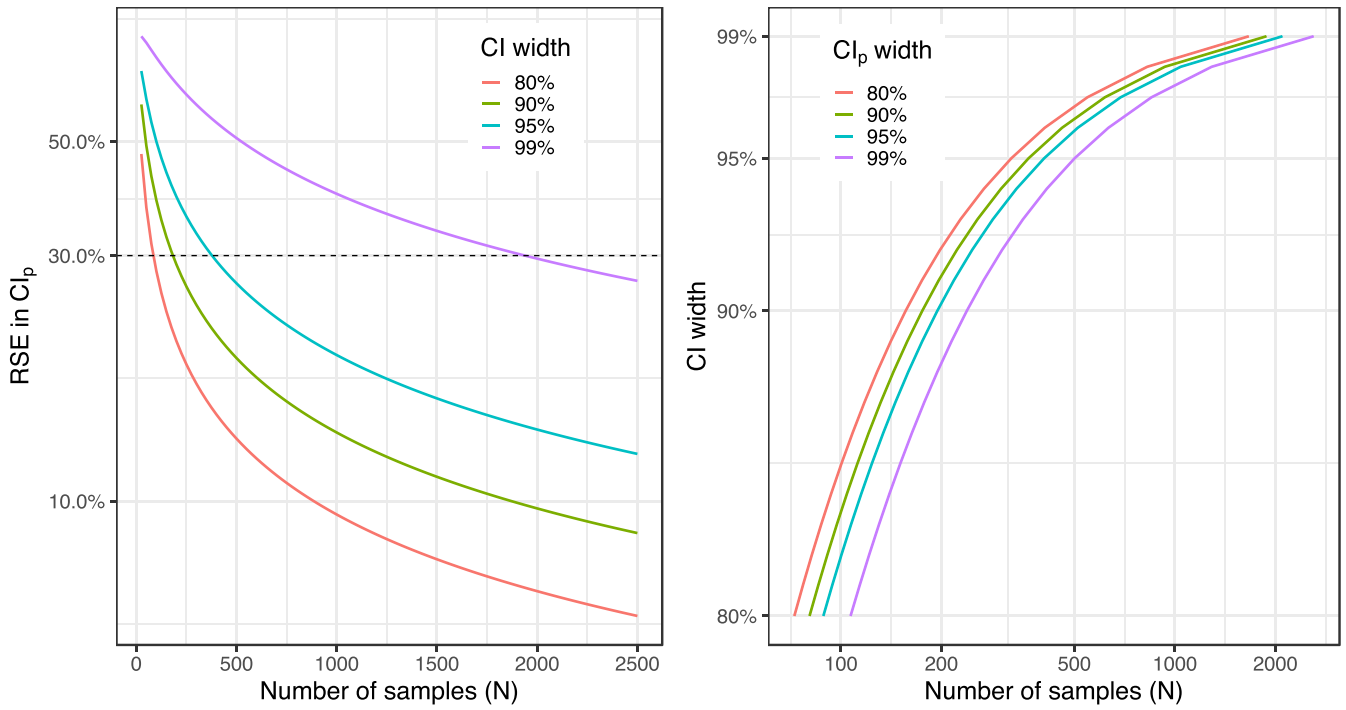

FIGURE 4 The left panel shows the RSE in CI_p versus the number of bootstrap samples. The colored lines correspond to a certain CI width and the dashed horizontal line indicates an RSE_p of 30% computed using Equation 2. The right panel shows CI width as a function of the number of bootstrap samples. The colored lines correspond to a certain CI_p width. CI, confidence interval; RSE, relative standard error

we can see that for a given CI width, the cost in terms of N increases the wider the desired CI_p is. Without the Agresti-Coull correction the four lines would have been superposed.

How precise do we have to be?

Equation 2 express the relationship between N, p, and RSE_p and describes that, for a given N, the smaller the p

is the larger the RSE_p becomes. In other words, the three factors involved are the desired width of CI, the desired degree of certainty in the estimation of the percentile limits and the number of bootstrap samples. All three are controlled by the analyst and can be adjusted to find an acceptable balance between inferential precision (the CI width), certainty in the inferential conclusions (the width of CI_p) and runtimes (the number of bootstrap samples, i.e., N).

The CI_p width is the uncertainty in the estimated uncertainty (i.e., CI) of the quantity of interest (e.g., a parameter estimate). Because an increasing CI_p width will increase the required N and because the CI_p width is not the primary interest (the CI is), it seems reasonable to choose a moderate value for the CI_p width. In the remainder of this paper, we will therefore base the calculations on a CI_p width of 90%.

The choice of N boils down to the question of how precise we need our estimated CI limits to be. Some analysis questions, such as "Is there a significant drug effect?" probably warrants a larger N than confidence intervals used in a Forest plot to illustrate the potential impact of covariates. One strategy is to use as many N as time permits and then choose a CI width that gives an acceptable precision in the CI limits. Another strategy is to choose a CI width and the level of precision in the CI limits we want and then derive the corresponding N.

In our organization, we have heuristically agreed on the rule that an acceptable precision in the CI limits is a 90% CI_p width of length $1-CI_{up}$ (i.e., the CI_p will cover 50% of the distance between CI_{lo}/CI_{up} and 0/1, respectively). For example, if the CI width is 95% and, consequently, the CI_{lo} and CI_{up} are 2.5% and 97.5%, respectively, then the target CI_p width should be 1–0.975 = 2.5% with the corresponding $CI_{p,lo}$ and $CI_{p,up}$ intervals of 1.25–3.75% and 96.25–98.75%, respectively.

With a rule like this, it is straightforward to calculate the desired N for different CI widths, as has been done in Table 1. It is obvious that wider CIs are disproportionately more expensive in terms of N compared to narrower CIs. Included in Table 1 is also the corresponding RSE in CI_{lo} and CI_{up}. Interestingly, it appears as if the heuristic rule used in the calculations results in RSEs of around 30–31%, except for the 0% CI width, which get an RSE of 35% with $N = 5$. However, as pointed out by Brown et al.,[10] the Agresti-Coull correction may not perform optimally for $N < 40$.

A CI of 95% would require 365 bootstrap samples to meet the criteria described above, whereas only 45 bootstrap samples are needed for a 68% CI (i.e., ±1 SD/SE). A 75% CI, the interquartile range, requires 62 samples. The CI width of 0% corresponds to a confidence interval around the midpoint in the bootstrap distribution (i.e., the median, requires five bootstrap samples according to these calculations but, as mentioned above, should probably be interpreted carefully given the small N).

TABLE 1 The number of samples required for different interval CI widths assuming the target $CI_{p,up}$ width is $1-CI_{up}$

CI width (%)	N[a]	RSE_p (%)[b]
99	1880	30
95	365	30
90	175	31
80	81	31
75	62	31
68	45	31
0	5	35

Abbreviations: CI, confidence interval; RSE, relative standard error.

[a]The number of required samples. Computed using the function calcN2 in the Supplementary Material S1.

[b]The relative standard error computed according to Equation 2 and refers to the uncertainty in $CI_{lo/up}$.

With Table 1 at hand it is possible to make an informed choice on the number of bootstrap samples to use. Using a default CI width of 90%, seems to represent a reasonable trade-off between precision and cost and is suitable for many applications, for example, in Forest plots. However, should more precise results be needed, it is possible to assess the cost in terms of, for example, run-times and evaluate if this is fit for purpose and/or predict when the results from the bootstrap analysis will be available.

Using a small bootstrap as an alternative to a nonparametric bootstrap for generating uncertainty estimates

Not directly related to percentile estimation, but still a pragmatic alternative in case a nonparametric bootstrap is too time-consuming, is to use the bootstrap to estimate only the standard errors of the primary or derived secondary parameters. This is done by running the bootstrap as usual and then take the standard deviation of the bootstrap estimates as the standard error of the bootstrap estimate of the parameter (i.e., the mean of the estimates from each bootstrap sample). The benefit is that a substantially lower number of bootstrap samples is required to obtain sufficient precision in this type of standard error estimate compared to estimating the limits of a bootstrap confidence interval. The drawback is that any confidence interval computed using this method will be symmetric. The theory behind the method is described by Ahn and Fessler[12] and shows that the RSE in an RSE estimate from

a bootstrap (bootstrap SE/bootstrap estimate) is given by Equation 7, provided the number of bootstrap samples is >10.

$$\text{RSE} = \frac{1}{\sqrt{2(N-1)}} \quad (7)$$

With Equation 7, it is easy to derive the expected RSE for an RSE estimated with a small bootstrap as just described. For example, the RSE in the estimated RSE for 10, 20, and 30 bootstrap samples would be 24%, 16%, and 13%, respectively. The estimated precision is much higher than for the limits of a nonsymmetrical confidence interval estimated with a nonparametric bootstrap (Table 1).

THE VISUAL PREDICTIVE CHECK

The VPC is used as a goodness of fit and general model checking diagnostic.[2] The principle is quite simple. The model is used to simulate a number of data sets (VPC samples) according to the same design as the original data set. Percentiles of the observed data across discrete points or bins of the dependent variable (usually time) are computed and plotted versus the independent variable. Typical percentiles are the 50th (the median), the 5th and 95th. The median line is a structural model (i.e., the fixed effects part of the model), diagnostic, whereas the outer percentiles are informative for the variability components of the model. The same percentiles as is derived from the observed data are then computed for each simulated data set and are visualized as intervals overlaying the percentiles from the observed data.

The simulated data used to illustrate the VPC below were generated using a single dose (=100 units), one compartment model with first order absorption, with the typical values of clearance, the volume of distribution, and the first order absorption rate constant set to 10, 100, and 2, respectively. Each of the parameters were associated with exponential interindividual variability (IIV) and an exponential residual unexplained variability (RUV) was used. The default value for the IIVs and RUV were ~30% and 15%, respectively. For illustration purposes, other values for IIV and RUV were also used (see below). The data were simulated with a very rich sampling design (40 samples per subject at identical timepoints across subjects, between 0 and 12 h with denser sampling around the maximum concentration) to support the visual presentation of the VPCs. No binning of observations was made (i.e., the VPC percentile summaries were done for each discrete timepoint).

Figure 5 shows a VPC using the example described above. The red lines are based on the observed data and the shaded areas around the red lines are based on the simulated data. In the context of VPCs, there are two types of intervals, the *prediction intervals* (i.e., the interval given by the difference between the outer percentiles of the observed data), and the *confidence intervals* (i.e., the shaded bands overlayed on the percentiles of the observed data), representing the confidence intervals around the percentiles of the simulated data.[2] However, this naming convention may not be entirely appropriate, see the Discussion. Both of these intervals are based on percentile estimation, similar to the nonparametric bootstrap, and are estimated with varying degrees of precision given by the number of subjects in the data set and the number of simulated data sets. It is worth pointing out that "simulating from the model," as is done in a VPC analysis, is in fact the same as doing a parametric bootstrap[13] but without the re-estimation of the model on each sample, and that we can use the same notation and principles as for the nonparametric bootstrap above. In this tutorial, it is assumed that each subject contributes one observation per independent value on the *x*-axis. With more or less observations per subject per independent value, the situation becomes more complicated but the general principles described herein still applies. Although there are similarities in the assessment of the precision in the prediction and confidence intervals in a VPC, there are also differences. The two types of intervals will therefore be discussed separately, starting with the prediction interval.

The VPC prediction interval

The width of the prediction interval (the distance between the dashed red lines in the VPC figures) in a VPC is determined by the variability in the data. This is illustrated in Figure 5, which is based on the simulations described above with alternative combinations of IIV (0%, 15%, and 30%) and RUV (0% and 15%) values. The larger the IIV and RUV, the wider the prediction interval. In the top left panel where the IIV and RUV are both set to 0%, the median and the prediction intervals overlap completely.

That the width of the prediction interval is independent of the number of subjects and number of VPC samples is illustrated in Figure 6, where the default model parameters were used to simulate data sets of different sizes (columns) and different number of VPC samples (rows). Because the red lines are based on the observed data and not the simulated data sets (VPC samples), they are identical in the panels with the same number of subjects (columns) and consequently so are the width of the prediction intervals. On the other hand, increasing the number of subjects in the observed data (left to right) reduce

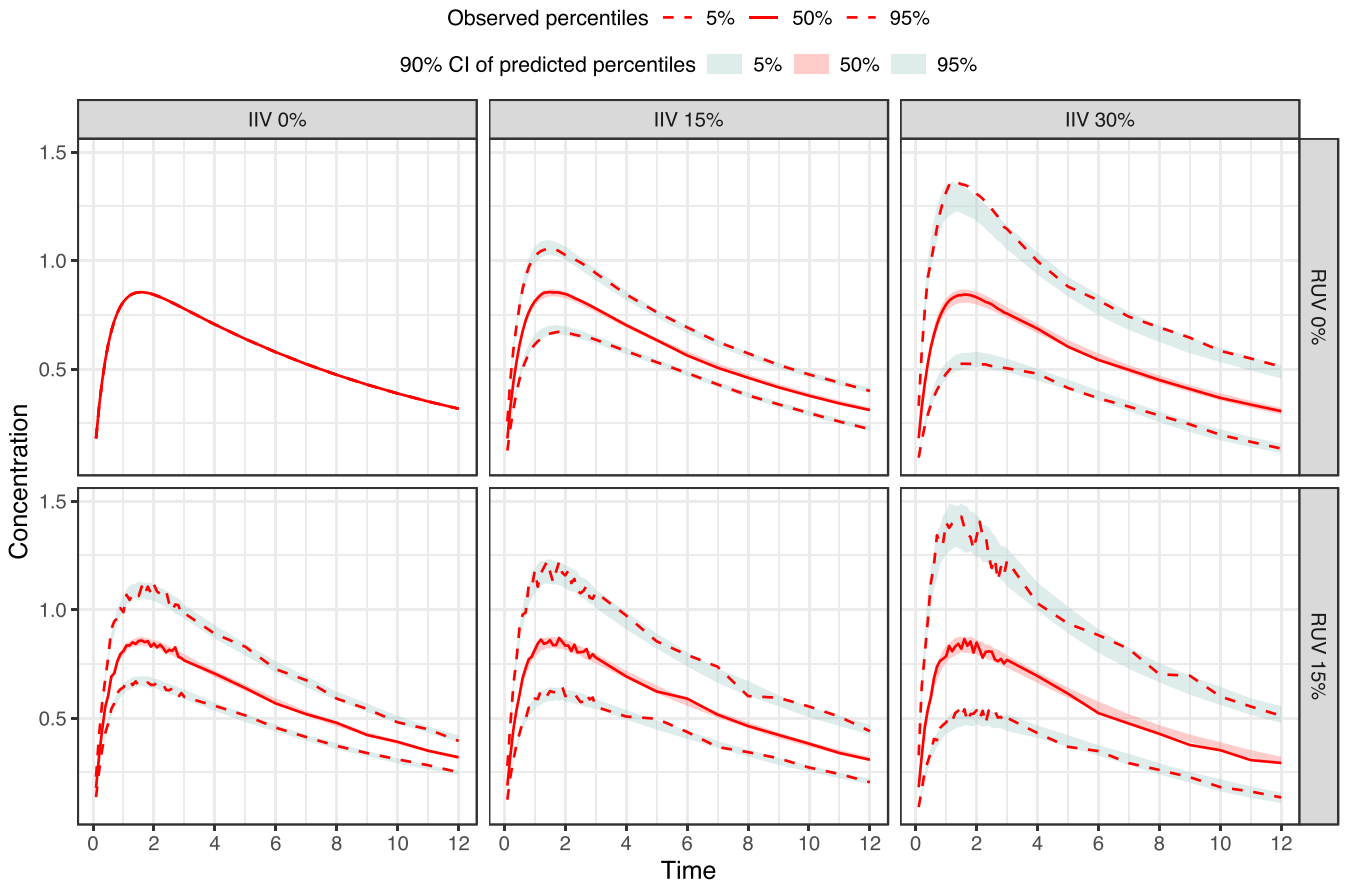

FIGURE 5 Schematic visual predictive check plots of a single dose concentration versus time profile. Each panel is based on 200 subjects simulated 200 times with varying IIV and RUV as given by the row and column titles. The red lines are generated from the observed data. The solid red line is the median of the observed data and the dashed red lines indicates the prediction interval. The shaded areas are the estimated confidence intervals (CIs) for the percentiles based on the simulated data. IIV, interindividual variability; RUV, residual unexplained variability

the variability in the red lines. This is exactly what would be expected from Equation 1—increasing N (number of subjects), given a certain prediction interval ($p = [1 - \text{prediction interval width}]/2$) will lead to a smaller SE.

How precise do we have to be?

Similar to the nonparametric bootstrap, the question is how precise we need to be in the estimation of the limits in a VPC's prediction interval?

It is hard to imagine a situation where it is necessary to be more precise in the estimation of the prediction interval in a VPC than what is required for the estimation of the CI limits in a nonparametric bootstrap, and whatever criteria we have for the bootstrap we can probably use for the prediction interval in a VPC. However, a difference in the VPC is that we cannot choose N for the prediction interval because it is given by the number of subjects. Instead, we need to adjust the prediction interval percentiles to obtain the desired precision.

Applying the heuristic rule described for the bootstrap on the prediction interval, we can use Table 1 to find the prediction interval width that is supported by different data set sizes (replacing CI with the prediction interval width). For example, 1000 subjects would support a 95% prediction interval with the desired precision and, similarly, 200 subjects would support a 90% prediction interval. In other words, the precision requirements from Table 1 for the 90% prediction interval used in Figure 6 are met by the 200 and 1000 subject data sets (middle and right columns) but not for the panels with 25 subjects, where the imprecision is larger than we want.

In many cases, it will be possible to adjust the prediction interval width according to Table 1 to obtain the target precision in prediction interval limits. For example, with a data set size of 175 subjects or more, a 90% prediction interval is supported, whereas 45 subjects support a 68% (one standard deviation) interval. Even though it is possible to derive (smaller) widths for smaller data set sizes, it is probably of little use because the percentiles are mainly a diagnostic for the variability components of the model,

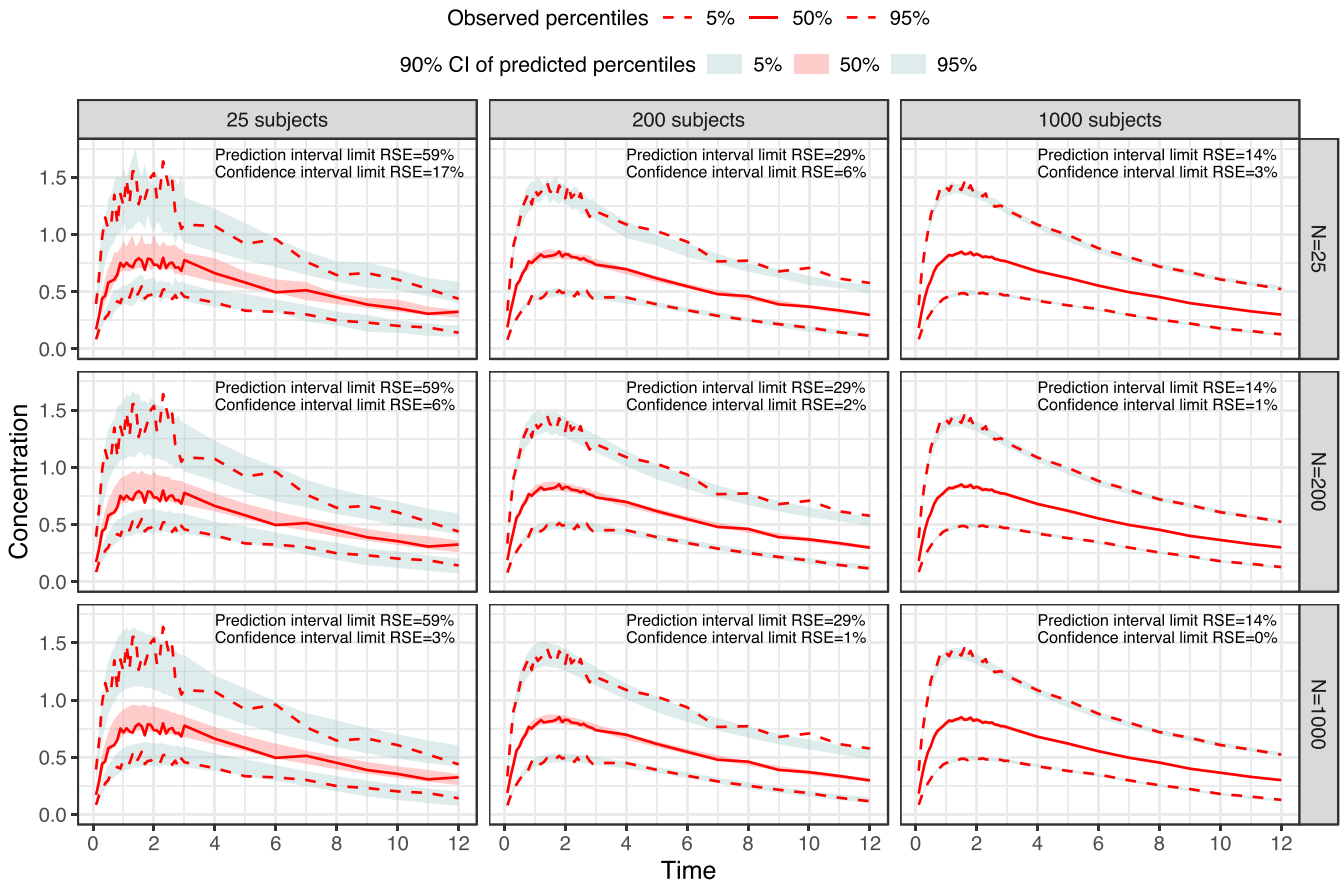

FIGURE 6 Schematic visual predictive check plots of a single dose concentration versus time profile. The data is generated using the default parameter values (30% IIV and 15% RUV). The number of subjects and number of samples used in each panel is given in the column and row titles The red lines are generated from the observed data. The solid red line is the median of the observed data and the dashed red lines indicates the prediction interval. The shaded areas are the estimated confidence intervals for the percentiles based on the simulated data. CI, confidence interval; IIV, interindividual variability; RSE, relative standard error; RUV, residual unexplained variability

and the fewer subjects the less information about the variability. However, even a small data set can be informative for the median line (a CI width of 0% in Table 1), which is useful as a structural model diagnostic.

In a practical situation where we may want VPCs stratified by covariates, we will likely have to handle different number of subjects in different strata. Figure 7 shows a VPC for our simulated example stratified by the covariate SEX. The number of male and female subjects are quite similar (95 vs. 105) and a prediction interval width of 80% has been chosen to match these numbers. In Figure 8, where the VPCs is stratified on the covariate study, it is only in the panel with study 1 among the top row panels, that the precision criteria from Table 1 is met for the 80% prediction interval that is used. This situation can be handled in different ways. One is to use different prediction interval widths in each panel, but that is likely to be quite confusing to the reader. Another approach is to use a prediction interval width that matches the strata with the smallest number of subjects (this is the recommended approach). In Figure 8 (top row), the situation is complicated by the small number of subjects in some of the studies, in particular study 5 with only six subjects. The solution, as mentioned above, is to not consider the prediction interval at all but to instead focus on the median. Figure 8 (bottom row) shows the same VPC as in Figure 8 (top row) but without the prediction interval. The relevant precision for the observed data in this plot is the uncertainties in the median lines, and they meet the criteria in Table 1. A display like this can preferably be combined with a prediction corrected VPC without stratification, or with an alternative stratification in which the strata are large enough to support the outer percentiles. However, another possibility is to make the *conscious* decision to still visualize the outer percentiles so that the appropriateness of the variability components of the model can be assessed in the strata with a higher number of subjects. In this case, the strata with the smaller number of subjects should not be over interpreted. The impact of stratification and binning in VPCs are discussed further below.

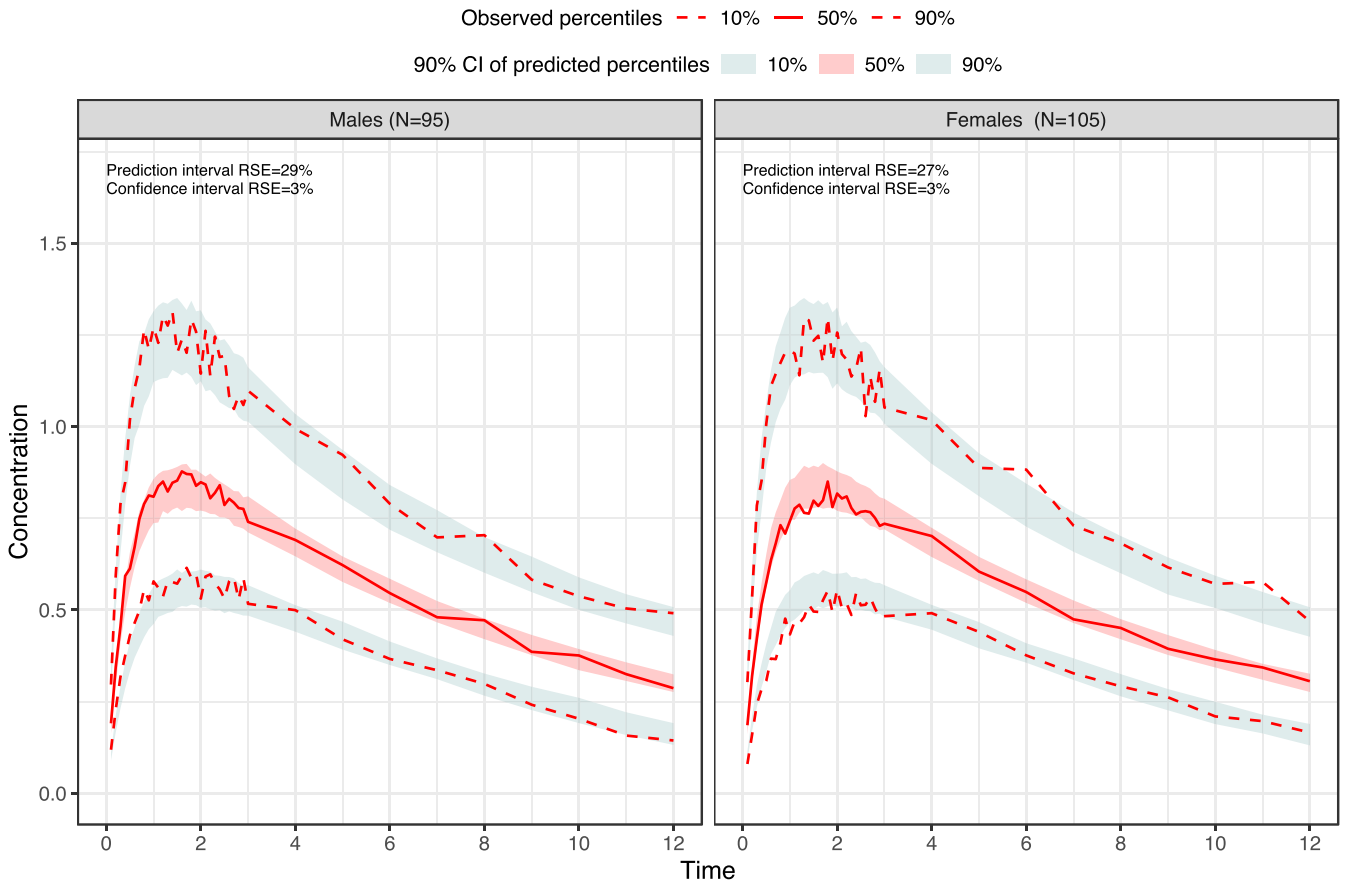

FIGURE 7 Visual predictive check plots based on the default example, stratified on the covariate Sex. The number of subjects in each panel is given in the panel title and each subject was simulated 200 times. The prediction interval widths are 80% and the confidence interval widths for the percentiles based on the simulated data are 90%. RSE, relative standard error

The VPC confidence interval

So far, we have focused on the VPC prediction intervals (the red lines), which are only dependent on the observed data. The confidence intervals in a VPC (the shaded bands overlaying the red lines) are generated by simulating several data sets from the model given the design of the observed data. In contrast to the nonparametric bootstrap, where CI_{lo} and CI_{up} is known without error and where the CI_ps therefore only depend on the number of bootstrap samples, the precision in the confidence interval limits in a VPC (which are analogous to CI_{lo} and CI_{up}) are also dependent on the uncertainty due to the number of subjects. The width of the confidence intervals in a VPC is given by the combination of the variability due to the number of subjects and the variability due to the number of VPC samples (Equation 8).

$$SD_{CI} = \sqrt{\frac{p_{PI}(1-p_{PI})}{N_{subj}} + \frac{p_{PI}(1-p_{PI})}{N_{samples}}} \quad (8)$$

SD_{CI} is the standard deviation of the percentile estimates from the simulated data sets, p_{PI} is 1 – the prediction interval/2 and N_{subj} and $N_{samples}$ are the number of subjects and the number of VPC samples, respectively.

The precision in the confidence interval limits, on the other hand, is related to the total number of simulated data points for each value (or bin) or the independent variable (Equation 9).

$$SE_{CI,limit} = \sqrt{\frac{p_{CI}(1-p_{CI})}{N_{samples} \times N_{subj}}} \quad (9)$$

$$RSE_{CI,limit} = \frac{\sqrt{\frac{p_{CI}(1-p_{CI})}{N_{samples} \times N_{subj}}}}{p_{CI}} \quad (10)$$

$SE_{CI,limit}$ is the standard error of the confidence interval limit, and p_{CI} is 1 – the confidence interval/2.

This means that there is a minimum width of the confidence interval which is given by the number of subjects (Equation 8) and this minimum width is independent of the number of VPC samples. This can be seen in Figure 6. Going from the top row to the bottom in the figure (i.e., keeping the number of subjects constant while increasing

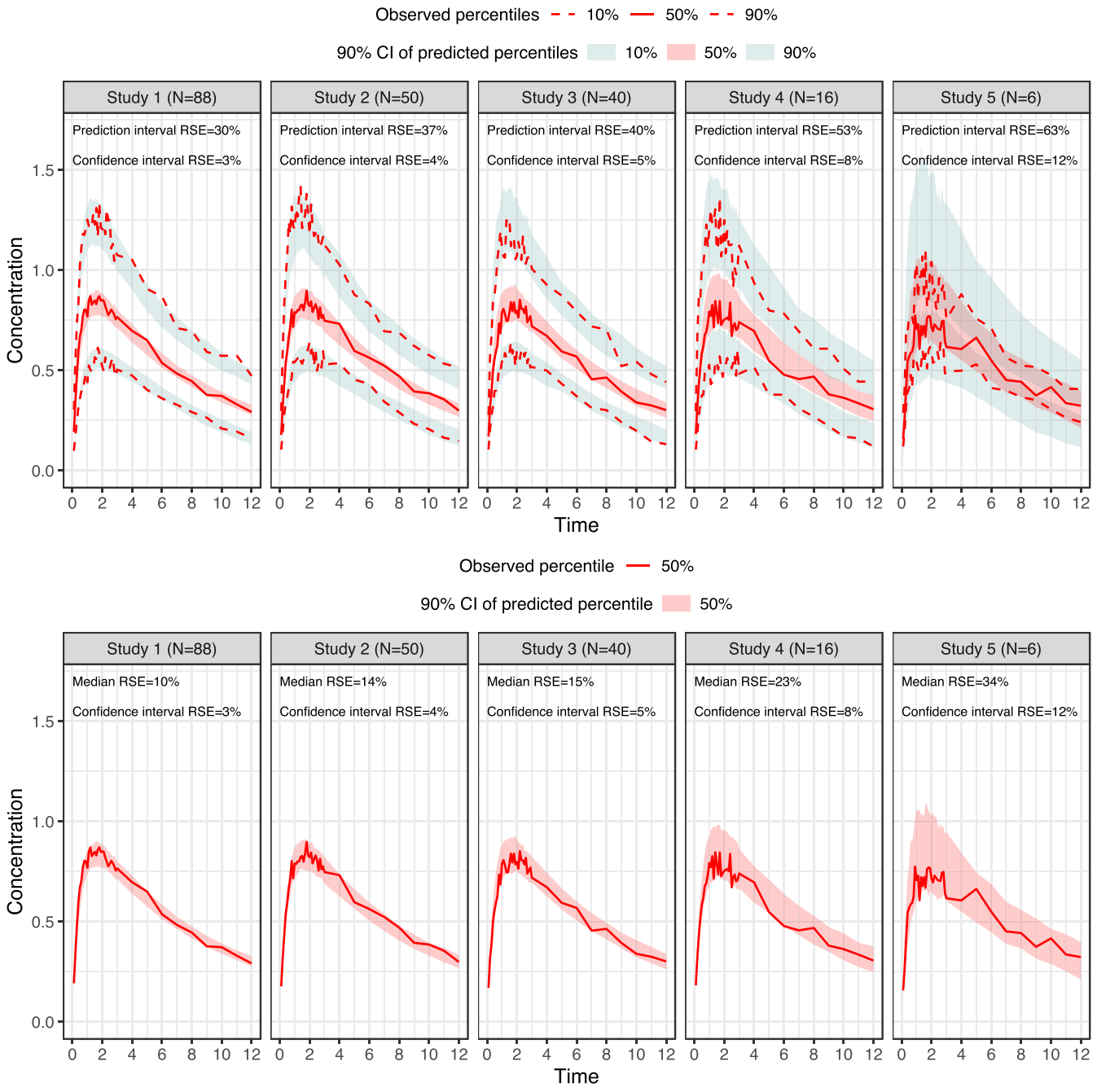

FIGURE 8 Visual predictive check plots based on the default example, stratified on the covariate Study. The number of subjects in each panel is given in the panel title and each subject was simulated 200 times. In the top row, the prediction interval widths are 80% whereas only the median is shown in the in the bottom row. The confidence interval widths for the percentiles based on the simulated data in both rows are 90%. CI, confidence interval; RSE, relative standard error

the number of VPC samples), the width of the confidence interval stays the same. On the other hand, going from left to right in Figure 6 (i.e., increasing the number of subjects while keeping the number of simulated data sets the same), decreases the width of the confidence interval in accordance with Equation 8.

Increasing the number of simulated data sets while keeping the number of subjects the same will decrease the uncertainty in the limits of the confidence intervals (Equation 9) (i.e., will generate smoother confidence intervals). This can be seen in, for example, the left column in Figure 6. In the top left panel, there is $25 \times 25 = 625$ observations for each timepoint, whereas there are $200 \times 25 = 2000$ and $1000 \times 25 = 25{,}000$ observations for each timepoint in the middle left and bottom left panels, respectively.

How precise do we have to be?

Because the number of subjects will be the main determining factor for the uncertainty in a VPC's confidence interval limits in most cases, there is no need to use a large number of VPC samples. Table 2 lists the RSEs for different combinations of number of subjects and number of VPC samples. The certainty in the estimation of the confidence interval limits very quickly becomes very high as the number of VPC samples increase.

If Table 1 is used to determine the width of the prediction interval (i.e., ~30% RSE), then it may be reasonable to require that the RSE in the confidence interval is lower than 30%. However, because the impact of uncertainty in the confidence limits is fairly limited (Figure 6) and because VPCs are not used for formal inference, it should not be necessary to use more than 200 simulated data sets (targeting an RSE < 10% for a data set/strata of 10 subjects). It should be acknowledged, though, that the assumption behind Table 2 is that there is one observation from each subject in the data set for each unique value of the independent variable or each bin in case the independent variable is binned, which should be kept in mind in case the actual situation is substantially different.

Impact of stratification and binning

In reality, it is likely that the VPCs are either stratified and/or use bins of the independent variable. The impact on the choice of prediction interval when one or more of the strata include only a few subjects has been discussed above. This also has implications for the number of VPC samples (Table 2), but the impact is lower because the precision is quite high already for small data sets and low sample numbers. However, binning adds an additional level of complexity. It may well be that not all bins contain observations from all subjects, or include only a handful of observation from only a few subjects. This means that some bins *within* a VPC plot may be associated with different degrees of certainty, both in the prediction interval limits, as well as in the confidence interval limits. Imbalances of this kind are hard to avoid but is worth paying attention to, for example, by choosing prediction interval widths and bin borders carefully so that too large precision differences are avoided.

DISCUSSION

Numerical percentile estimation as done in the nonparametric bootstrap and the VPCs are associated with uncertainty. In this tutorial, we have provided a quantitative framework for assessing the precision in estimated percentile limits in bootstrap and VPC analyses, to facilitate an informed choice among confidence interval width, number of VPC/bootstrap samples, and required level of precision.

It is worth pointing out that a very precise bootstrap estimate of uncertainty is not the same as that the estimated uncertainty is a good reflection of the true uncertainty. In the paper by Dosne et al.,[14] they showed that the bootstrap was unsuitable for datasets including up to 70 individuals but concluded that the number of subjects by itself is not sufficient as a predictor of bootstrap appropriateness. Other aspects, such as individual study design, model misspecifications, overparameterization, bootstrap stratification strategies, and handling of boundary

TABLE 2 The RSE (according to Equation 10) in a 90% VPC confidence interval across different data set sizes and samples

N_{subj}[a]	$N_{samp_{25}}$[b] (%)	$N_{samp_{50}}$ (%)	$N_{samp_{100}}$ (%)	$N_{samp_{200}}$ (%)	$N_{samp_{500}}$ (%)	$N_{samp_{1000}}$ (%)
10	27.6	19.5	13.8	9.7	6.2	4.4
25	17.4	12.3	8.7	6.2	3.9	2.8
50	12.3	8.7	6.2	4.4	2.8	1.9
75	10.1	7.1	5	3.6	2.3	1.6
100	8.7	6.2	4.4	3.1	1.9	1.4
125	7.8	5.5	3.9	2.8	1.7	1.2
150	7.1	5	3.6	2.5	1.6	1.1
175	6.6	4.7	3.3	2.3	1.5	1
200	6.2	4.4	3.1	2.2	1.4	1
500	3.9	2.8	1.9	1.4	0.9	0.6
1000	2.8	1.9	1.4	1	0.6	0.4

Abbreviations: RSE, relative standard error; VPC, visual predictive check.

[a]The number of subjects in the data set.

[b]The number of samples, as indicated by the subscript.

conditions, in the bootstrap re-estimations also influence the appropriateness of the bootstrap results. However, in this tutorial, which focuses on how well the uncertainty is estimated and not how good the uncertainty estimate is, it is assumed that the concerns raised by Dosne et al.[14] are appropriately handled.

The VPCs in this tutorial do not include any parameter uncertainty (but can be added using the $PRIOR functionality in NONMEM, for example). Including the parameter uncertainty will lead to wider confidence intervals (i.e., making it easier for a bad model to look "good" in the VPC analysis), but would not affect the uncertainty in the confidence interval limits. We believe that the performance of a model should always be illustrated without the parameter uncertainty included. If there is a need to motivate why an apparent model misspecification is inside the possibilities of the posterior parameter distribution, this can be illustrated with a second VPC in which the uncertainty is included or using other methods in which the parameter uncertainty is accounted for.

In the case of the bootstrap, the precision in the uncertainty can be controlled by the user through the number of bootstrap samples and the width of the confidence interval. In the case of the VPC, the amount of observed data determines the precision in the prediction interval limits (the observed data) and, together with the number of VPC samples, the precision in the confidence interval limits (the simulated data). The combined influence of the number of subjects and the number of VPC samples, together with stratification and binning, makes it more complicated to control the confidence interval uncertainty in the VPC than in the bootstrap. On the other hand, the VPCs are not used for making the type of inferential decision that the bootstrap can be used for so the lower degree of control of the uncertainty is probably less of a problem. One important difference, however, is that in the bootstrap it is possible to use the number of bootstrap samples to minimize the uncertainty in the confidence interval limits while in the VPC, the minimum uncertainty is given by the number of observed data points.

In the bootstrap, given the run-time cost of adding bootstrap samples, the question is how many bootstrap samples that are required to meet a particular certainty level, whereas in the VPC the question is more about avoiding an unnecessary large number of VPC samples. Note that we have referred to the interval around the predicted percentiles in the VPC as "confidence interval" to be consistent with the original paper.[2] It is, however, debatable if this is strictly a confidence interval because the width of this interval in a VPC depends on both the amount of observed data (N_{subj} in Table 2) and the number of VPC samples and will not become narrower than what is given by the first term in Equation 8. The precision in the confidence interval width, on the other hand, will increase with the number of VPC samples regardless of the number of subjects.

The computational cost of the nonparametric bootstrap has been mentioned several times in this tutorial and is a practical limitation of the method. The longer the run time of the model, the more costly the bootstrap. Because bootstraps are usually carried out toward the end of an analysis it is important to choose the number of bootstrap samples, or width of the target confidence intervals, in such a way so that the delivery of the final results is not delayed. If the heuristic rule behind Table 1 is accepted, then 175 bootstrap samples for a 90% confidence interval is sufficient (or 365 for a 95% confidence interval, in case a higher degree of confidence is needed), which is substantially smaller than the 1000 that is often used.[4–7]

The computational cost of the VPC is often smaller than for the bootstrap, because it is purely based on simulations, but the VPC is commonly carried out more frequently during a pharmacometric analysis than the bootstrap. The VPC is a powerful overall goodness of fit assessment tool that can be used in most stages and, indeed, after most runs in a pharmacometric project. However, even if the VPC only involves simulations it can still take time, as can the required post-processing calculations to summarize the simulations into confidence intervals. Both the simulation and post-processing times are proportional to the number of simulated data sets (VPC samples) as well as to the size of the observed data, and may be prohibitive for frequent use if a large number of simulated data sets is used. Fortunately, it seems as if a large number of simulated data sets is rarely needed, especially if the number of subjects is high (Table 2). By carefully assessing the minimum number of data points in the bins in the smallest strata and adjusting the bin borders to avoid bins with only a few observations, it should be possible to select a sample size that makes it possible to use the VPC as standard goodness of fit instrument for all runs in a pharmacometric project. Specifically, it seems likely that less than 200 simulated data sets should be sufficient in most situations.

ACKNOWLEDGEMENTS

The authors thank Professor Mats O. Karlsson for valuable input, and Dr Siv Jönsson and Dr Martin Bergstrand for review of early drafts.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

ORCID

Joakim Nyberg https://orcid.org/0000-0002-2839-4940

REFERENCES

1. Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat*. 1979;7:1-26.
2. Bergstrand M, Hooker AC, Wallin JE, Karlsson MO. Prediction-corrected visual predictive checks for diagnosing nonlinear mixed-effects models. *AAPS J*. 2011;13:143-151.
3. Menon-Andersen D, Yu B, Madabushi R, et al. Essential pharmacokinetic information for drug dosage decisions: a concise visual presentation in the drug label. *Clin Pharmacol Ther*. 2011;90:471-474.
4. Dosne A-G, Bergstrand M, Karlsson MO. An automated sampling importance resampling procedure for estimating parameter uncertainty. *J Pharmacokinet Pharmacodyn*. 2017;44:509-520.
5. Dosne A-G, Bergstrand M, Harling K, Karlsson MO. Improving the estimation of parameter uncertainty distributions in nonlinear mixed effects models using sampling importance resampling. *J Pharmacokinet Pharmacodyn*. 2016;43:583-596.
6. Thai H-T, Mentré F, Holford NHG, Veyrat-Follet C, Comets E. A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. *Comparative Study*. 2013;12:129-140.
7. Thai HT, Mentré F, Holford NH, Veyrat-Follet C, Comets E. Evaluation of bootstrap methods for estimating uncertainty of parameters in nonlinear mixed-effects models: a simulation study in population pharmacokinetics. *J Pharmacokinet Pharmacodyn*. 2014;41:15-33.
8. Gastonguay MR, El-Tahtawy A. Effect of nonmem minimization status and number of replicates on bootstrap parameter distributions for population pharmacokinetic models: a case study. *Clin Pharmacol Therap*. 2005;77:P2.
9. Lomax R, Hahs-Vaughn D. *An Introduction to Statistical Concepts*. Taylor & Francis Group; 2013.
10. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Stat Sci*. 2001;16:101-117.
11. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. *Am Stat*. 1998;52:119-126.
12. Ahn S, Fessler JA. *Standard Errors of Mean, Variance, and Standard Deviation Estimators*. EECS Department, The University of Michigan; 2003:1-2.
13. Davison AC, Hinkley DV. *Bootstrap Methods and their Application*. Cambridge University Press; 1997.
14. Dosne AG, Niebecker R, Karlsson MO. dOFV distributions: a new diagnostic for the adequacy of parameter uncertainty in nonlinear mixed-effects models applied to the bootstrap. *J Pharmacokinet Pharmacodyn*. 2016;43:597-608.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Jonsson EN, Nyberg J. A quantitative approach to the choice of number of samples for percentile estimation in bootstrap and visual predictive check analyses. *CPT Pharmacometrics Syst Pharmacol*. 2022;11:673-686. doi:10.1002/psp4.12790