J Pharmacokinet Pharmacodyn (2008) 35:185–202
DOI 10.1007/s10928-007-9081-1

Extensions to the Visual Predictive Check to facilitate model performance evaluation

Teun M. Post · Jan I. Freijer · Bart A. Ploeger · Meindert Danhof

Received: 6 August 2007 / Accepted: 5 December 2007 / Published online: 16 January 2008

Abstract The Visual Predictive Check (VPC) is a valuable and supportive instrument for evaluating model performance. However in its most commonly applied form, the method largely depends on a subjective comparison of the distribution of the simulated data with the observed data, without explicitly quantifying and relating the information in both. In recent adaptations to the VPC this drawback is taken into consideration by presenting the observed and predicted data as percentiles. In addition, in some of these adaptations the uncertainty in the predictions is represented visually. However, it is not assessed whether the expected random distribution of the observations around the predicted median trend is realised in relation to the number of observations. Moreover the influence of and the information residing in missing data at each time point is not taken into consideration. Therefore, in this investigation the VPC is extended with two methods to support a less subjective and thereby more adequate evaluation of model performance: (i) the Quantified Visual Predictive Check (QVPC) and (ii) the Bootstrap Visual Predictive Check (BVPC). The QVPC presents the distribution of the observations as a percentage, thus regardless the density of the data, above and below the predicted median at each time point, while also visualising the percentage of unavailable data. The BVPC weighs the predicted median against the 5th, 50th and

T. M. Post
NV Organon, Oss, The Netherlands

B. A. Ploeger · M. Danhof (✉)
Leiden/Amsterdam Center for Drug Research, Division of Pharmacology, Leiden University,
P.O. Box 9502, 2300 RA, Leiden, The Netherlands
e-mail: m.danhof@lacdr.leidenuniv.nl

B. A. Ploeger · M. Danhof
Leiden Experts on Advanced Pharmacokinetics & Pharmacodynamics, Leiden, The Netherlands

J. I. Freijer
Astellas Pharma, Leiderdorp, The Netherlands

95th percentiles resulting from a bootstrap of the observed data median at each time point, while accounting for the number and the theoretical position of unavailable data. The proposed extensions to the VPC are illustrated by a pharmacokinetic simulation example and applied to a pharmacodynamic disease progression example.

Keywords Visual Predictive Check · Pharmacokinetics · Pharmacodynamics · Disease progression · Evaluation · Validation · Qualification · Model performance

Introduction

Throughout the various phases in drug development, data are increasingly analysed using population pharmacokinetic (PK), pharmacodynamic (PD) and disease progression (DP) models [1–4]. An important part in the modelling process is to evaluate how well a model performs, either to gain knowledge on how to improve the model from a learning perspective or to confirm its validity for further application, e.g. for clinical trial simulations [5]. Various methods exist for this purpose, many of which are graphical or statistical evaluations of the observations in relation to measures of the model prediction [6]. An appealing method for checking model performance is the so-called Visual Predictive Check (VPC), which originates from the Posterior Predictive Check (PPC) [7,8]. The VPC is a valuable tool as it is simply applicable from a practical standpoint and has the potential for strong intuitive convincing, albeit subjective, graphs [7,9]. The rationale of the VPC is to graphically assess whether an identified model is able to reproduce the variability in the observed data from which it originates. Using the identified model, several replications based on the structure of the original data (i.e. dosing, timing and number of samples) are simulated and the distribution of these replicates is compared to actual observations. Figure 1 shows the most common display of the VPC where the 5th and 95th percentiles, and in some cases the 50th percentile, of the simulated distribution are compared to the observations. In this general form, the VPC largely depends on a subjective comparison of the overlap in the simulated distribution with the observations, without explicitly quantifying and relating the information in both. Recently proposed adaptations to the VPC methodology clearly take this relation into account by showing the percentiles of the distribution of observed and simulated data [10–16]. Visual evaluation of the resemblance between the observed and the predicted variability and the general time trend in the data, while also including the anticipated but unavailable data, is an elementary feature in model evaluation [8,17,18]. Although the VPC is currently a supportive tool, it has important shortcomings in its present application, which limits its use as a standard tool for model evaluation [9].

In general, three main shortcomings in the standard VPC can be identified. These shortcomings are partly accounted for in the recently published adaptations, but require further advances, which are described in this paper: (i) it is not evaluated whether the median calculated from the simulations is well-matched to the median of the observations, (ii) neither the distribution of the observations, nor the expected random allocation of the observations around the model predicted median are objectively considered and (iii) neither the amount of observations at each time point, nor the influence and

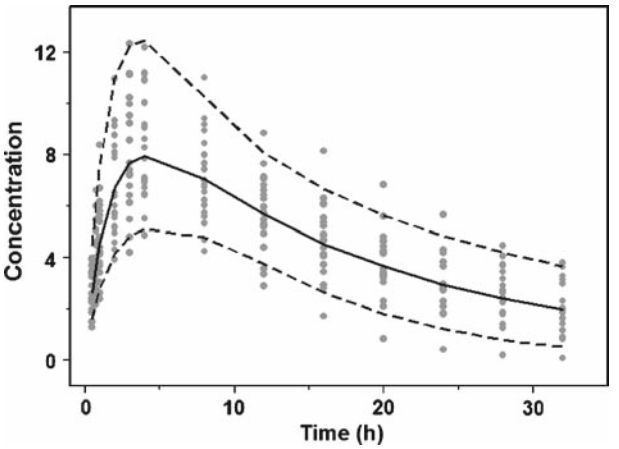

Fig. 1 Common display of the VPC. The dark grey dots present the observations, the dashed black lines the 5th and 95th percentiles of the model simulation and the solid black line depicts the model predicted median. The observations arise from the PK simulation MODEL 1 and the VPC from optimisation MODEL 2 (Table 1)

information residing in unavailable data are taken into account in relation to either the first or the second issue. Within the context of this paper, the term 'unavailable data' reflects all sorts of missing or unobserved data resulting from a variety of reasons, e.g. concentrations below the lower limit of quantification, subject drop-out, random or informative missing. As the aim of the manuscript is the clarification of the conceptual extensions to the VPC, it is judged reasonable to use this term throughout the manuscript.

Specific VPC examples exist where observations and model predictions are graphically linked. Winter et al. and Holford et al. display the median of the observations next to the model predicted median, managing the first shortcoming [10,11]. Furthermore, others have presented examples in which the observed data are displayed as percentiles, presenting the distribution of the observed data around the model predicted median, thereby also accommodating the second shortcoming [11,13–15]. Following submission of this paper, R-scripts have become available which take this one step further by enabling the visualisation of uncertainty ranges around the percentiles of the observed and simulated data [15]. Recently, another method based on the PPC was proposed, which deals with the first two issues [16,19]. This attractive approach computes the prediction discrepancies for each observation and summarises the position of the observations in the distribution of predictions graphically and statistically. These recent developments clearly show the evolution in methodology and tools for performing a VPC aiming to overcome the shortcomings of its elementary form. Nevertheless, the aforementioned approaches do not account for the amount of unavailable data and the information residing in these data, thereby disregarding their possible influence on the observed data distribution and, indirectly the model predictions. This introduces an inequality in the comparison between the model prediction and the observations.

In order to address the identified deficiencies two extensions are proposed; the Quantified Visual Predictive Check (QVPC) and the Bootstrap Visual Predictive Check (BVPC). In general, the first extension visualises the distribution of the available data around the model predicted median while including the amount of unavailable data at each time point. The second extension displays measures of the distribution of the bootstrapped observations, while accounting for the influence (number and theoretical position) of unavailable data. This reflects the existing uncertainty in the observed data used for model identification and renders a more objective comparison with the predicted median. As the amount of observations will influence the visual identification of an adequate model, the uncertainty in the observations from which the model is derived should be clearly shown. In this respect, the QVPC shows the data relative to the model and the BVPC the model relative to the data.

This paper describes two extensions to the VPC in a general framework, enabling a more accurate and objective evaluation of model performance. To illustrate the characteristics and the application of these methods two examples are presented. The first is a PK simulation example in which the characteristics of the methods are exemplified. The second example concerns the application of the methods in PD disease progression modelling.

Methods

Quantified Visual Predictive Check (QVPC)

As the median is defined as the value in the middle of the observation range (i.e. 50% of the values are greater or lower than this respective value), it is expected that for an adequate model the observations at each time point are randomly allocated around the model predicted median. The QVPC is a method that enables the visualisation of the positioning of available observations regardless of the density, the amount and the shape of the data distribution in relation to the model predicted median in a VPC. Unavailable observations are also taken into account as they can have an influence on the graphical interpretation of the VPC, for example when data are missing on one side of the distribution with non-random dropout [11,20] or when data are below the limit of quantification. The predicted median is selected, as opposed to the mean, as it is not sensitive to outliers and it presents the best measure of central tendency if the distribution is skewed.

Let $n_{aobs,t}$ denote the number of available observations at time t, M_t the model predicted median corresponding to the 50th percentile of the simulated distribution at time t and N the expected number of observations at each time point given the number of individuals included in the study. The percentage of available observations around M_t and the percentage of unavailable observations at each time point are then reflected by:

$$100\% = \begin{cases} A_{M,t}(\%) = \sum (n_{aobs,t} \geq M_t)/N \cdot 100 \\ B_{M,t}(\%) = \sum (n_{aobs,t} \leq M_t)/N \cdot 100 \\ U_{M,t}(\%) = 100 - (A_{M,t} + B_{M,t}) \end{cases} \quad (1)$$

where $A_{M,t}$ is the percentage of available observations above and $B_{M,t}$ the percentage of available observations below the model predicted median at time t and $U_{M,t}$ presents the percentage of unavailable observations at time t.

The median of the observed data m at time t is reflected by 50% if the available number of observations at each time point ($n_{aobs,t}$) equal the expected number of observations (N):

$$m_t = 50\% = ((A_{M,t} + B_{M,t})/2) + U_{M,t} \quad \text{with } U_{M,t} = 0\% \tag{2}$$

In the theoretical situation of a perfect model prediction with all observations available ($U_{M,t} = 0\%$ and $A_{M,t} = B_{M,t} = 50\%$,), m_t equals both $A_{M,t}$ and $B_{M,t}$. When all observations are available but the model prediction is inadequate at certain time points, m_t will still equal 50%. However, $A_{M,t}$ and $B_{M,t}$ will contrariwise differ from 50%, totalling 100%. In this manner a poor distribution around the model predicted median is visualised, which helps identifying model misspecification. In case part of the data are unavailable ($U_{M,t} > 0\%$), m_t will deviate from 50% revealing the weight of unavailable observations on the actual median of the observed data at specific time points. Larger deviations of m_t from 50% point to the need of a more subtle interpretation of the model prediction and the observed data distribution around it, as less data exists for the interpretation of model performance. Further clarification and interpretation of the QVPC characteristics are presented in the results section.

Bootstrap Visual Predictive Check (BVPC)

The QVPC examines whether the observations are randomly scattered around the median prediction. However, the percentages above and below the model predicted median will rarely be exactly 50% at each time point as the observed data on which the model was obtained is uncertain depending on the density, amount and shape of its distribution and the influence of the unavailable data. This uncertainty in the median of the observed data should be characterised before comparing the observed to the model predicted median. The BVPC presents a method that identifies and visualises this uncertainty by computing a nonparametric bootstrapped median including its 5th and 95th percentiles based on the available data per time point, while accounting for the number and theoretical position of unavailable data. The code was written in S-PLUS and is attached to the paper as appendix.

Let $y_{aobs,t}$ denote an available observation, $n_{aobs,t}$ the number of available observations at time t and N the expected number of observations at each time point given the number of individuals that were included in the study. In the BVPC the following algorithm is performed for each time point:

1. Draw a number of bootstrap replications (*NoBS*), e.g. 1000 times, from $y_{aobs,t}$ with replacement. Each sample of the 1000 replications now has a number of observations that equals $n_{aobs,t}$.
2. Compose a new dataset (*BootSamp*) with the number of rows equal to $n_{aobs,t}$ and the number of columns equal to *NoBS*.

3. Fill each position in *BootSamp* with the possible realisations obtained in step 1 of the observed data from the statistics of the bootstrap.

When all observations are available at each time point ($U_{m,t} = 0\%$), $n_{aobs,t}$ equals the expected number of observations (N) and *BootSamp* automatically contains as many rows as N. The median of each replicate (*NoBS*) in *BootSamp* can be obtained directly and the 5th, 50th and 95th percentiles of these replicate medians are determined. These percentiles are included in the VPC graph and reflect the observed data median and its uncertainty limits and will be referred to as *bootstrapped median*.

In the situation where data are unavailable at certain time points ($U_{m,t} > 0\%$), $n_{aobs,t}$ does not equal the expected number of observations (N) and *BootSamp* contains as many rows as $n_{aobs,t}$. The difference between N and $n_{aobs,t}$ presents the portion of the unavailable data at these time points ($n_{uobs,t}$). The possible influence of these data on the *bootstrapped median*, depending both the amount and position of these data, should be considered in the evaluation. Therefore, the following additional steps are performed:

4. Obtain $n_{uobs,t}$ and find the minimum and maximum values of $y_{aobs,t}$.
5. Compose a second dataset (*UnavSamp*) with the number of rows equal to $n_{uobs,t}$ and the number of columns equal to *NoBS*.
6. Fill each position in *UnavSamp* with either the minimum or maximum value of $y_{aobs,t}$ obtained in step 4 given an assumed probability of whether it is located at the higher or lower end of the observed data distribution. For a random allocation, when no influence of unavailable data on the median of the observed data is assumed, the probability is set to 50%.
7. Merge *BootSamp* and *UnavSamp* thereby ensuring that the combined dataset (*Both*) contains as many rows as N.

By taking $n_{aobs,t}$ and $n_{uobs,t}$ jointly into account, the sample size in the replications as obtained in step 3 will now correctly equal the expected amount of data (N) at each time point before calculating the *bootstrapped median*. In addition to the amount of unavailable data, the position of the unavailable data in the data distribution is also accounted for. For example, when data are below the limit of quantification, it is expected that a number of unavailable observations are located below the observed data median and these are then assigned this position within the data distribution.

Finally, the *bootstrapped median* is calculated and included in the VPC graph, reflecting the observed data median and its uncertainty limits. The percentiles are determined only when the fraction of unavailable observations account for less than 50% of the expected observations at a certain time point. Otherwise the percentiles are considered inaccurate as the uncertainty in the observations has such an influence that evaluation of model performance is not realistic at that time point.

Application QVPC and BVPC

Pharmacokinetic example (simulated data)

Pharmacokinetic profiles were simulated for 20 subjects receiving a 100 mg oral dose of a hypothetical compound. A total of 13 samples at 0.5, 0.75, 1, 2, 3, 4, 8, 12,

Table 1 Parameters of the population PK simulation and optimisation models

Parameter (units)	MODEL 1 Simulation Value	MODEL 2 All data Optimised (%CV)	MODEL 3 Data above LOQ 3 Optimised (%CV)	MODEL 4 Data above LOQ 5 Optimised (%CV)
ka (h^{-1})	0.6	0.609 (2.32)	0.642 (2.73)	0.674 (5.95)
V (l)	10	9.79 (7.08)	10.1 (7.08)	9.93 (6.95)
CL (lh^{-1})	0.5	0.523 (6.94)	0.497 (6.70)	0.450 (10.0)
IIV[a] CL (%)	30	30.5 (27.1)	27.9 (32.2)	31.6 (44.0)
IIV[a] V (%)	30	30.3 (25.5)	29.7 (25.3)	26.1 (22.3)
Residual[b] (%)	10	8.94 (10.9)	8.37 (13.3)	8.37 (15.3)

[a] Inter-individual variability presented as $\sqrt{\omega^2} \cdot 100\%$
[b] Proportional residual error presented as $\sqrt{\sigma^2} \cdot 100\%$

16, 20, 24, 28 and 32 h were collected. Nominal times were used in order to present the methods more clearly. A one compartment model with first-order absorption and first-order elimination was used, parameterised in terms of volume of distribution (V) and clearance (CL). Inter-individual variability on V and CL was described using a log-normal distribution. A proportional error model was used to describe the residual error. Subsequently, the identical PK model structure as used in the simulations was optimised on three different datasets: (i) the full dataset, (ii) a dataset with data below 3 concentration units removed and (iii) a dataset with data below 5 concentration units removed. A model building step was not performed and values below the quantification limit were not replaced by, for example half the quantification limit, in order to compare the results solely based on the VPC extensions. The identified parameter estimates of the three models were then used to simulate a 1000 replicates of the data based on the study design. With these simulations, the corresponding VPC and the extensions QVPC and BVPC were obtained. Table 1 presents the parameter values of the simulation and optimisation models. In the BVPC all unavailable data were assumed to be located at the lower end of the observed data distribution.

Pharmacodynamic disease progression example (actual phase III data)

To demonstrate the application of the methods on a real dataset, they were applied on a phase III Type 2 Diabetes Mellitus (T2DM) study using a previously established mechanism-based disease progression model [10]. The HbA_{1c} data of in total 1204 T2DM patients treated with pioglitazone during a 1-year period were examined using the described methods. As the observations were randomly collected within dedicated time periods, nominal times were assigned to represent these intervals. These nominal times were 1, 15, 43, 71, 99, 127, 183, 239, 309 and 379 days. Initially in the BVPC, all unavailable data were assumed to be randomly located around the model predicted median.

Computation

The simulations and optimisations were performed by non-linear mixed-effects modelling using NONMEM version V, release 1.1 (Icon Development Solutions, Ellicott City, MD, USA) on desktop computers with an AMD Athlon processor under Windows 2000. The models were compiled using Digital Fortran (version 6.6, Compaq Computer Corporation, Houston, TX). Optimisations were performed with the FOCE with interaction method. The presented methods were developed using the statistical software package S-PLUS for Windows (version 6.2 Professional, Insightful Corp., Seattle, WA, USA).

Results

Pharmacokinetic example (simulated data)

Figure 1 displays the simulated concentration vs. time data arising from PK simulation MODEL 1 and the respective VPC resulting from optimisation MODEL 2 (Table 1). The dark grey dots present the observations, the dashed black lines the 5th and 95th percentiles of the model simulation and the solid black line depicts the model predicted median (M). This figure, based on the full dataset, reflects the standard VPC format as a starting point from which the methods are exemplified. The QVPC and BVPC plots of this example are presented in Fig. 2 and the parameter estimates in Table 1. Upon removal of data throughout the example, the point estimate of clearance progressively decreases and the uncertainty in this parameter estimate increases as a result of an increased uncertainty arising from a smaller amount of data.

All data

The QVPC corresponding to the VPC based on the full dataset is presented in Fig. 2; upper row. As all expected concentrations are available in this dataset both bars reflecting the percentage above ($A_{M,t}$: dark grey) and below ($B_{M,t}$: black) the model predicted median (M_t) approximate 50%. The boundary between the two bars presents the position of M_t within the observed data. Furthermore, the reflection of the observed data median (m_t: white dots) equals 50% at each time point as all data are available. In the ideal situation where all expected data are available, the model fit is perfect and a substantial amount of subjects are included ($m_t = A_{M,t} = B_{M,t} = 50\%$), the boundary between $A_{M,t}$ and $B_{M,t}$ should equal m_t. However, the dataset consists of a relatively small number of 20 subjects, thereby introducing a certain amount of uncertainty in median of the observations on which the model was optimised. The corresponding BVPC reflects this uncertainty as is evident from Fig. 2. The dark grey area depicts the range between the 5th and 95th percentiles of the *bootstrapped median*, clearly characterising the uncertainty range in the median of the observations. The model predicted median is allowed to vary within this uncertainty range. The dashed white line symbolises the 50th percentile of the *bootstrapped median*, which presents

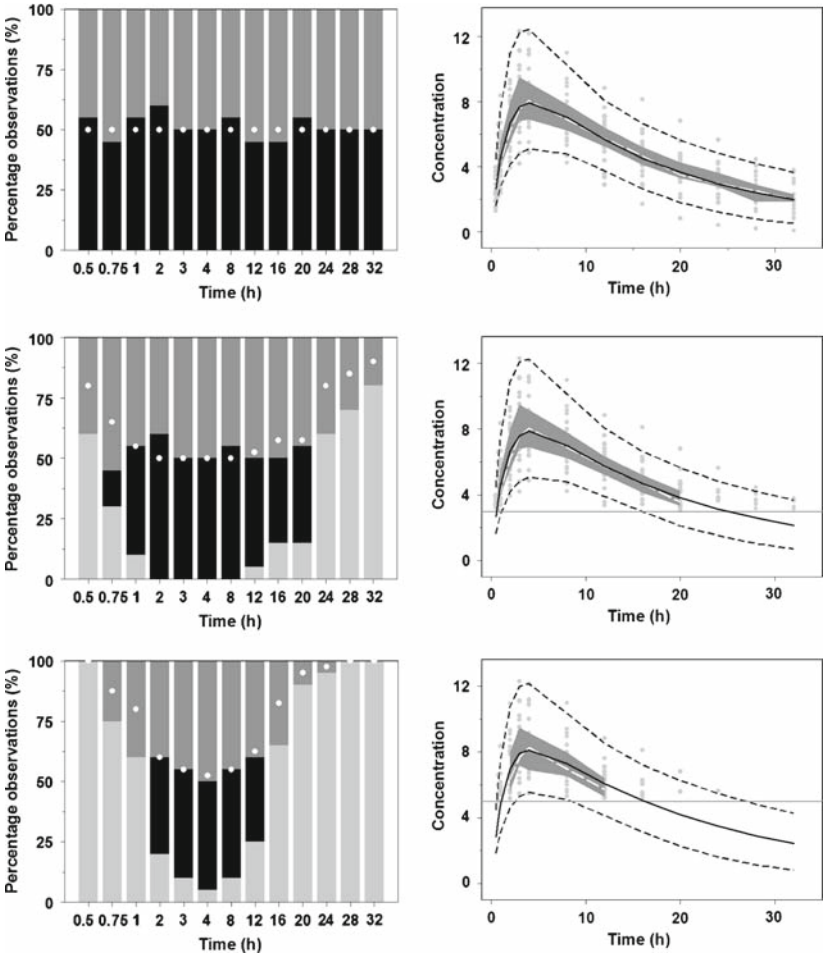

Fig. 2 Quantified Visual Predictive Check (QVPC; left) and Bootstrap Visual Predictive Check (BVPC; right). Upper row: full dataset population PK model (Table 1, MODEL 2). Middle row: observations below 3 concentration units were removed (Table 1, MODEL 3). Lower row: observations below 5 concentration units were removed (Table 1, MODEL 4). QVPC: The distribution of the observed data around the model predicted median at each observation time (M_t) is presented as a percentage of the expected amount of data. The black bar presents the observed data below M_t, the dark grey bar the observed data above M_t. The total of the black and grey bar combined presents the percentage of available data ($n_{aobs,t}$). The light grey bar presents the percentage of unavailable observations. The white dots symbolize the percentage of $n_{aobs,t}$ divided by 2 on top of the percentage of unavailable observations and reflect the observed data median. BVPC: The light grey dots present the available observations. The dashed black lines present the 5th and 95th percentiles and the solid black line depicts the model predicted median of the standard VPC. The dark grey area depicts the range between the 5th and 95th percentiles of the *bootstrapped median*, which reflects the uncertainty range in the median of the observations. The dashed white line represents the 50th percentile of the *bootstrapped median*. For the models where the unavailable observations are removed, the probability for these data was set to be 100% below the model predicted median

an additional link between the model predicted trend and that in the observed data. Both QVPC and BVPC present an adequate fit of the model to the data.

Data above LOQ 3

The QVPC and BVPC based on the model optimisation on the dataset where data below 3 concentration units are removed are presented in Fig. 2; middle row. Table 1 presents the parameter estimates of MODEL 3 based on this reduced dataset. The percentage unavailable observations ($U_{m,t}$) are visualised in the QVPC by light grey bars in addition to the bars reflecting the percentage above ($A_{M,t}$) and below ($B_{M,t}$) the model predicted median (M_t). The graph shows that between 2 and 8 h all expected data are available and both $A_{M,t}$ and $B_{M,t}$ as well as the observed data median (m_t) approximate 50%. Therefore, the model can be readily evaluated at these time points. At remaining time points, the unavailable data influence the interpretation of the model prediction as shown by m_t which deviates from 50%. The evaluation of model performance should then be more subtle. The graph shows that most data are above the model predicted median, which in this simulation example obviously relates to the fact that data below 3 concentration units were removed. For example, between 8 and 20 h, the percentage of data above the model predicted median approximates 50%, whereas that below the model predicted median is smaller. In addition, the observed data median parallels the boundary between the two areas at these time points, indicating an adequate model prediction.

In this example, the uncertainty in the median of the observations is further attenuated by the presence of unavailable data. In the BVPC graph, this is presented by the range (dark grey area) between the 5th and 95th percentiles of the *bootstrapped median* and the related 50th percentile (dashed white line). The amount and theoretical position of the unavailable observations are accounted for in the *bootstrapped median*, where the location of these data is understood to always be below the model predicted median. The model predicted median (solid black line) lies well within the uncertainty range and follows the trend of the 50th percentile of the *bootstrapped median*. The QVPC previously indicated an adequate model performance between 8 and 20 h and the BVPC confirms this finding. At 0.5 h and after 20 h the percentage of unavailable observations is greater than 50% and the measures of the *bootstrapped median* are omitted as they cannot be accurately determined. This informs that at these time points an unambiguous evaluation of model performance is not feasible.

Data above LOQ 5

The QVPC and BVPC based on the model optimisation on the dataset where data below 5 concentration units are removed are presented in Fig. 2; lower row. Table 1 shows the parameter estimates of MODEL 4 based on this reduced dataset.

The QVPC shows that at 4 h almost all expected data are available and that the allocation around the model predicted median is adequate at this time point ($A_{M,t} \approx B_{M,t} \approx m_t$), although slightly more data are above the model predicted median (M_t). At 3 and 8 h some data are unavailable but $A_{M,t}$ approximates 50%. Nevertheless, slightly more data are below M_t when taking into account the unavailable data.

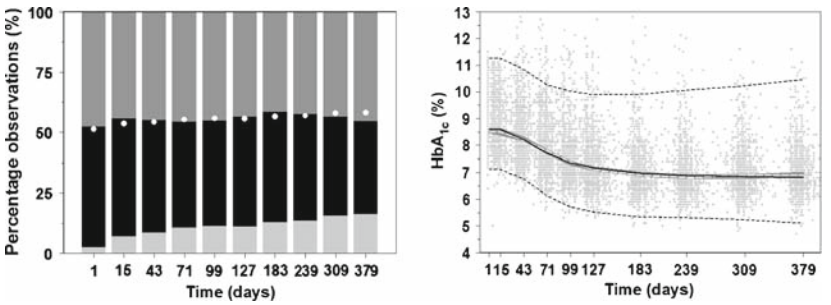

Fig. 3 Quantified Visual Predictive Check (QVPC; left) and Bootstrap Visual Predictive Check (BVPC; right) of the mechanism-based disease progression model on HbA_{1c} measurements. Legends as described for Fig. 2. In the BVPC a probability of 50% is set for the unavailable data in the *bootstrapped median* assuming a random allocation of the unavailable observations around the model predicted median

At 2 and 12 h $A_{M,t}$ appears to be structurally smaller than 50%, indicating an inferior model estimation. The BVPC further supports this view as M_t moves away from the 50th percentile of the *bootstrapped median* at these time points. Even so, M_t lies within the range of the *bootstrapped median*, indicative of a sufficient model fit at these time points given the uncertainty in the observations. At the remaining time points the *bootstrapped median* cannot be accurately determined and an evaluation of model performance is not feasible.

Pharmacodynamic disease progression example (actual phase III data)

The results of the QVPC and BVPC of the mechanism-based disease progression model [10] performed on the phase III HbA_{1c} data are presented in Fig. 3.

The QVPC shows a cumulative amount of unavailable data over time as would be expected in a long-term trial due to drop-out. As a result the reflection of the observed data median (m_t) deviates from 50% over time. Although $A_{M,t}$ and $B_{M,t}$ are not 50% they are approximately equal, presenting a random allocation of the available observations around M_t. As there are a large number of observations and the model predicted median (boundary between $A_{M,t}$ and $B_{M,t}$) follows m_t, this gives an indication that the model prediction is adequate. At 379 days somewhat more data are located above the model predicted median (M_t), indicating a slight underprediction of the trend at this time point.

In the BVPC the amount of unavailable observations is taken into account and the allocation of these data is assumed to be random. Due to the substantial amount of observations the uncertainty in the observed data median is small and consequently the range of the *bootstrapped median* is narrow. Notice that the distribution of the *bootstrapped median* is skewed, thereby properly reflecting the distribution of the actual data. The model predicted median is generally well-matched to the 50th percentile of the *bootstrapped median*, indicating an adequate model performance. Nevertheless, the underprediction of the trend as observed in the QVPC is also apparent in the BVPC and the model predicted median lies just outside the narrow uncertainty range of the *bootstrapped median*.

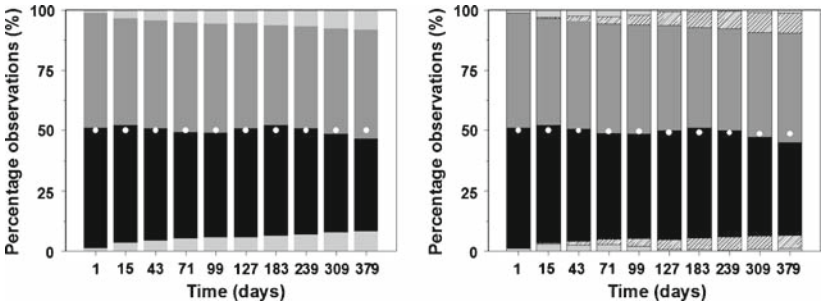

Fig. 4 Quantified Visual Predictive Check (QVPC) of the mechanism-based disease progression model on HbA_{1c} measurements. Legends as described for Fig. 2. The QVPC on the left has the percentage of unavailable observations randomly allocated around the model predicted median (light grey bar). The QVPC on the right has a pattern bar included which presents the location of the unavailable observations around the model predicted median for patients leaving the study based on the position of the last available observation relative to the model predicted median (pattern bar). The light grey bar now presents the percentage of remaining unavailable observations randomly allocated around the model predicted median

In order to further examine the most likely position of the unavailable observations, they are subsequently assumed to be randomly allocated around M_t in the QVPC. This gives a first impression of whether inclusion of, for example a drop-out model should be attempted. Figure 4; left graph, presents a QVPC with the unavailable observations appointed evenly above and below M_t. Markedly, the reflection of the observed data median (m_t) approximates 50% when considering the unavailable data as being randomly missing, indicating that these data can be considered randomly allocated around M_t. Furthermore, the model predicted median (boundary between $A_{M,t}$ and $B_{M,t}$) follows m_t when considering the unavailable data randomly allocated, except for the previously mentioned time point where a slight underprediction of the trend was observed.

As a subsequent step, the nature of the total of unavailable observations is investigated. The amount and position of the unavailable observations that are structurally missing are separated from the observations that are infrequently missing. Figure 4; right graph, presents a QVPC graph where unavailable observations of patients leaving the study are assigned above or below the predicted median according to their position relative to the predicted median at the time point before drop-out, carrying this position forward towards the last time point in the study (pattern bar). For example, when a patient left at time t and the position of HbA_{1c} at time $t-1$ was above M_{t-1} this location is then reserved for the rest of the study duration. See for an analogous example the paper by Hu and Sale [20]. Half of the remaining observations that apparently are infrequently or randomly unavailable are presented above and the other half are presented below M_t. At 379 days more patients cumulatively left the study with a HbA_{1c} measurement above than below M_t, indicative of a possible influence of these subjects on the model predicted time trend, either on the treatment effect or on the parameters reflecting the total disease progression.

A further advantage of the methods becomes apparent when examining model performance in an external validation when the confirmation dataset consists of a

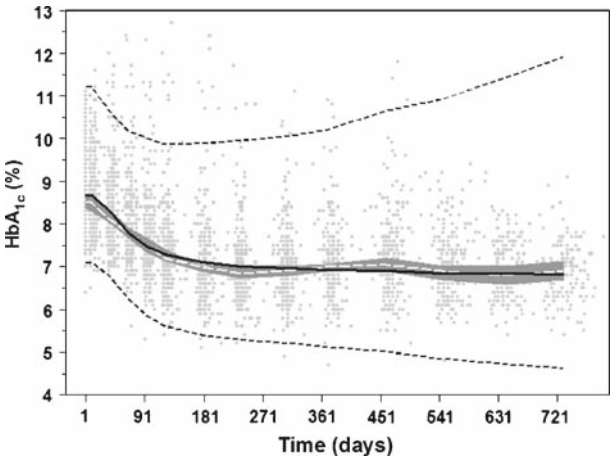

Fig. 5 Bootstrap Visual Predictive Check (BVPC) of the mechanism-based disease progression model optimised on 1-year data on HbA_{1c} measurements of 270 patients during a 2-year period. Legends as described for Fig. 2 with a probability of 50% set for the unavailable data in the *bootstrapped median* assuming a random allocation of the unavailable observations around the model predicted median

dissimilar amount of subjects than that on which the model was identified. The QVPC and BVPC then link the model prediction and the new observations more objectively, despite a varying quantity of data. An example of such a situation is presented in Fig. 5 where the BVPC of the mechanism-based disease progression model optimised on 1-year data is presented in relation to the HbA_{1c} observations of a subset of 270 patients that continued treatment for a second year [23]. During the second year of treatment, it appears that the variability is overestimated as the 5th and 95th percentiles of the VPC fan out and the observations are only within the lower part of the prediction interval. It could be argued that time trend in the data is predicted too optimistically by the model, focussing mainly on the observations of the patients that are still present in the study. However, as the uncertainty in the observations is taken into account in addition to the unavailable observations as presented by the *bootstrapped median,* a comparison with the model predicted median is feasible. Note also the increase in the uncertainty range as presented by the *bootstrapped median* with an increasing amount of unavailable data during the second year. From the BVPC it is clear that the two-year time trend on this subset of patients can be predicted adequately by the mechanism-based model which was optimised on 1-year data. Furthermore, the perceived underestimation of the time trend at the end of the first year now appears to be related more to the system than to the influence of dropout as the predicted time trend in the second year is quite accurate. As such, the methods thus substantially optimise the use of the available information when evaluating model performance.

Discussion

The described methods in this paper present practical and effective extensions to the VPC that render this model evaluation method more objective and accurate. The basis

of the proposed methods lies in the fact that the observed data are explicitly linked to the model predictions, and vice versa, while including the influence of the expected but unavailable data [8,17,18]. The benefit of the QVPC arises from the visual allocation of the available and unavailable observations around the model predicted median regardless of the density and the shape of the observed data distribution. The pertinent feature to account for the uncertainty in the observed median while accounting for the influence of missing data in relation to the model prediction is a clear advantage of the BVPC over the standard VPC.

Illustration QVPC and BVPC

The presented PK simulation example illustrates the structure and the interpretation of the conceptual methods in a situation where progressively more data are unavailable. This example shows which influence unavailable data exert on the evaluation of model performance using a VPC. The QVPC objectively relates the observations to the model prediction by quantifying the amount of observations that surround the model predicted median. Without this method speculation of the observed data distribution is standard practice, as currently done by interpreting Fig. 1. From this graph it is solely evaluated whether the prediction interval overlaps with the perceived distribution of the observed data. Furthermore, the interpretation of the QVPC is based on a "complete" dataset as the amount of unavailable data is displayed, which additionally reduces the speculation of the amount of data that are supposedly observed. The QVPC thus offers distinct advantages over a standard VPC, as in the latter the model might be judged as being inadequate despite the fact that it could merely be a result of interpreting only available data without consideration of the misleading distribution of the available data or the underlying influence of unavailable data.

Besides relating the observations to the model prediction in the QVPC, the model prediction is also related to the characteristics of the data, as presented by the BVPC. The inherent uncertainty in the observations on which the model was optimised are characterised with this method, thereby presenting a means to objectively position the model predicted trend within the observations. Ultimately, the example shows that the QVPC and BVPC as presented in Fig. 2 are in fact able to reveal an underestimation of clearance resulting from a substantial amount of observations below the quantification limit. This confirms the finding by Hing et al. who showed that omitting observations below the lower limit of quantification will result in an underestimation of clearance in specific cases [21]. Therefore, the extensions present a practical tool to diagnose whether these unobserved data have an impact. In the manner the VPC is currently applied this would not have been immediately apparent. For PK data most of the unavailable data, consciously disregarding actual randomly missing data, can readily be assigned to be missing at the lower end of the data distribution, making these extensions a blueprint for these kind of studies.

Application QVPC and BVPC

The PD disease progression example shows the application of the VPC extensions on a real phase III dataset. These trials primarily consist of large patient populations

where the amount of observations and the shape of the data distribution influence and complicate an objective judgment of the model performance. By application of the QVPC and BVPC, the substantial amount of 1204 observed and unobserved data per time point are objectively displayed, offering a comprehensive insight into the data and model prediction. The methods show an apparent underestimated time trend, as displayed in Fig. 3. The standard VPC would not be able to detect this as 1204 observations in relation to the model prediction cannot be "eyeballed". Furthermore, due to various reasons a cumulative amount of data are unavailable in long term trials, obscuring the true data distribution that would have been present with all data available. Both extensions to the VPC offer improved diagnostic tools in such a setting.

Moreover, the added benefit of the QVPC is apparent when unavailable data are examined further. In these particular trials the number of expected but unobserved data is related to either patients prematurely leaving the study or to randomly missing data, which both distort the evaluation of model performance [11,14]. The influence on the reflection of the observed data median can be visualised when randomly allocating the missing data above and below the model prediction. Figure 4; left graph displays the HbA_{1c} data in this manner and the reflection of the observed data median now closely resembles 50%. Figure 4; right graph then presents a further division of the unavailable data into random and nonrandom missing data. For the patients prematurely leaving the study the last available observation is examined for its position relative to the model predicted median and this location is projected forwards. The remaining unavailable observations are allocated randomly around the model predicted median. Now the distribution and allocation of all unavailable observations can be observed according to their origin of absence and this presents valuable information. From Fig. 4; right graph it becomes apparent an increasing number of patients that discontinue treatment are located above the model predicted median. This presents a possible explanation for the identified underestimated time trend although other reasons such as adaptation of the system or the onset of the treatment effect could present alternative explanations. Nevertheless, it enables the acknowledgement of the possible influence of subject drop-out before formally addressing it in a modelling process, which can be a complex and time consuming process [17,20,22].

Combined, the methods result in simple, yet effective graphical diagnostics that ensure that: (i) it is evaluated whether the model predicted median is well-matched to the observed data median and its uncertainty, (ii) the distribution of the observations and the expected random allocation of the observations around the model predicted median are considered in an objective manner and (iii) the amount of observations at each time point and the influence and information residing in unavailable data are taken into account in relation to the first and second point.

In general, the VPC can readily be used in trials with a fixed sampling schedule or a fixed time window of sampling. However, in the situation where PK or PD measurements are obtained at varying time points among subjects or when dense and sparse profiles are both available within a study, the methods are more complex, as the identification of a representative median or the allocation around the model predicted median per time point are more challenging to identify. A solution for this would be to substitute the number of expected subjects per study, as was assumed in this paper, with the expected number of subjects per time point. Another issue arises in the

situation where measurements are obtained at random time points within dedicated time windows and subjects have multiple measurements within one time window due to additional visits related to safety evaluation or subject drop-out. This should be accounted for as this would artificially increase the amount of available data at that time interval. Nevertheless, these issues are certainly disregarded without a proper link between the observations and the model prediction in the manner the VPC is currently applied.

Furthermore, the application of the VPC is limited by the fact that a more or less fixed dosing schedule or substantial subsets of individuals with identical dosing schedules are required to generate comprehensive VPC plots in which data of identical groups can be combined and shown in isolation that present a substantial body of information. When the data arising from these groups becomes too little, the VPC and the proposed extensions to the VPC have limited use as the amount of information becomes too small to meaningfully visualise the data or to calculate the *bootstrapped median*. This requirement of attaining substantial subsets to be able to graphically display the VPC also holds true for trials in which various dose groups, sampling schedules or covariate levels are investigated. Finally, the VPC can be only practical in the situation where observations on a continuous scale are obtained. For trials investigating pharmacodynamics on a categorical scale the VPC and the proposed extensions are of limited use.

As the purpose of this paper was to introduce and clarify the conceptual extensions to the VPC, it was not investigated whether they are able to discriminate between competing models as was pursued by Jadhav and Gobburu [9]. However, the improvement in diagnostics will support the evaluation of competing models by taking into account all aspects of the models and data.

In summary, two conceptual methods are presented with dedicated examples illustrating simple, yet effective extension to the VPC to evaluate model performance. Both methods result in a more objective and accurate interpretation of the VPC by explicitly linking the observed data and the model prediction, while accounting for the influence of expected but unavailable data.

Appendix

```
MyBootDrop <- function(x,probupdrop,nodropout)
# Call MyBootDrop for each timepoint in the dataset
# out <- MyBootDrop(x_t, P, n_uobs,t)
# out <- MyBootDrop(data[data$TIME==t, "DV"],0.5,n_uobs,t)
# x           - vector with observations per timepoint (y_aobs,t)
# probupdrop  - probability of being censored at the upper tail
#             - 1 = all at the upper end, 0 = all at the lower end, 0.5
#                = random
# nodropout   - number of unavailable observations per timepoint
#               (n_uobs,t)
# Result      - Bootstrapped median (with 95% uncertainty limits)
{
        # Number of boostrap samples
        NoBS = 1000
        # bootstrap dataset with replicates all saved in BSM
```

```
    BSM <- bootstrap(x, median(x),seed=5,B=NoBS,save.indices=T)
    # all bootstrap replicates of median
    Indices <- resamp.get.indices(BSM)
    # Determine extremes in dataset
    maxx <- max(x)
    minx <- min(x)
    # bootstrap samples
    BootSamp <- matrix(nrow=length(x),ncol=NoBS)
    BootSamp[,] <- x[Indices[,]]
    # Make matrix with extremes to account for unavailable data
    UnavSamp<-matrix(nrow=nodropout,ncol=NoBS)
    # Fill matrix with theoretical unavailable observations
    if (nodropout>=1)
    {
         for (j in 1:NoBS)
         {
                  for (i in 1:nodropout)
                  {
                      # probupdrop: probability of being allocated
                      # above (1), below (0) or random (0.5) around
                      # the model predicted median
                      yes<-rbinom(1,1,probupdrop)
                      UnavSamp[i,j]<-maxx*yes+minx*(1-yes)
                  }
         }
         # Bind extremes with BootSamp
         Both<-rbind(BootSamp,UnavSamp)
    }
    else
    {
         Both<-BootSamp
    }
    result<-rep(3,NoBS)
    # Determine medians of all replicate sets
    for (k in 1:NoBS)
    {
         result[k] <- median(Both[,k])
    }
    # Determine percentiles of the replicate medians
    if (nodropout>=length(x))
    {
         # if unavailable data account for more or equal the amount
         # of available observations the bootstrapped median is
              omitted
         QNT<-quantile(c(-99,-99,-99),c(0.05,0.5,0.95), na.rm = F)
    }
    else
    {
         QNT<-quantile(result,c(0.05,0.5,0.95), na.rm = F)
    }
    # Return result
    return (QNT)
}
```

References

1. Bhattaram AV et al (2007) Impact of pharmacometrics review on new drug approval and labeling decisions – a survey of 31 new drug applications. Clin Pharmacol Ther 81:213–221
2. Sheiner LB, Steimer JL (2000) Pharmacokinetic/pharmacodynamic modelling in drug development. Annu Rev Pharmacol Toxicol 40:67–96
3. Lalonde RL, Kowalski KG, Hutmacher MM, Ewy W, Nichols DJ, Milligan PA et al (2007) Model-based drug development. Clin Pharmacol Ther 82(1):21–32
4. Chan PL, Holford NH (2001) Drug treatment effects on disease progression. Annu Rev Pharmacol Toxicol 41:625–659
5. Sheiner LB (1997) Learning versus confirming in clinical drug development. Clin Pharmacol Ther 61(3):275–291
6. Brendel K, Dartois C, Comets E et al (2007) Are population pharmacokinetic and/or pharmacodynamic models adequately evaluated?: a survey of the literature from 2002 to 2004. Clin Pharmacokinet 46(3):221–234
7. Holford N (2005) The Visual Predictive Check – Superiority to Standard Diagnostic (Rorschach) Plots. PAGE 14, Abstr 738 [www.page-meeting.org/?abstract=972]
8. Yano Y, Beal SL, Sheiner LB (2001) Evaluating pharmacokinetic/pharmacodynamic models using the Posterior Predictive Check. J Pharmacokin Pharmacodynam 28(2):171–192
9. Jadhav PR, Gobburu JVS (2005) A new equivalence based metric for predictive check to qualify mixed-effects models. AAPS J 7(3):E523–E531
10. de Winter W, DeJongh J, Post T et al (2006) A mechanism-based disease model for comparison of long-term effects of pioglitazone, metformin and gliclazide on disease processes underlying type 2 diabetes mellitus. J Pharmacokin Pharmacodynam 33(3):313–343
11. Holford NHG, Chan PLS, Nutt JG, Kieburtz K, Shoulson I, Parkinson Study Group (2006) Disease progression and pharmacodynamics in Parkinson Disease – evidence for functional protection with levodopa and other treatments. J Pharmacokin Pharmacodynam 33:281–311
12. Post TM, Freijer JI, de Winter W et al (2006) Accurate interpretation of the visual predictive check in order to evaluate model performance PAGE 15, Abstr 972 [www.page-meeting.org/?abstract=972]
13. Holford N, Pillai G, Kaila N et al (2006) PKPD model for cathepsin K inhibition with balicatib and changes in bone turnover biomarkers, in particular NTx PAGE 15 Abstr 1015 [www.page-meeting.org/?abstract=1015]
14. Fang L, Holford NHG, Hinkle G et al (2007) Population pharmacokinetics of humanized monoclonal antibody HuCC49ΔCH2 and murine antibody CC49 in colorectal cancer patients. J Clin Pharmacol 47(2):227–237
15. Urien S, Holford N (2007) R for NONMEM. https://sourceforge.net/project/showfiles.php?group_id=29501&package_id=140129&release_id=538680
16. Mentré F, Escolano S (2006) Prediction discrepancies for the evaluation of nonlinear mixed-effects models. J Pharmacokin Pharmacodynam 33(3):345–367
17. Gelman A, Van Mechelen I, Verbeke G et al (2005) Multiple imputation for model checking: completed-data plots with missing and latent data. Biometrics 61:74–85
18. Sheiner LB, Beal SL, Dunne A (1997) Analysis of nonrandomly censored ordered categorical longitudinal data from analgesic trials. J Am Stat Assoc 92:1235–1255
19. Brendel K, Comets E, Laffont C, Laveille C, Mentré F (2006) Metrics for external model evaluation with an application to the population pharmacokinetics of gliclazide. Pharm Res 23(9):2036–2049
20. Hu C, Sale ME (2003) A joint model for nonlinear longitudinal data with informative dropout. J Pharmacokin Pharmacodynam 30(1):83–103
21. Hing JP, Woolfrey SG, Greenslade D, Wright PMC (2001) Analysis of toxicokinetic data using NONMEM: impact of quantification limit and replacement strategies for censored data. J Pharmacokin Pharmacodynam 28(5):465–479
22. Unnebrink K, Windeler J (2001) Intention-to-treat methods for dealing with missing values in clinical trials of progressively deteriorating diseases. Stat Med 20:3931–3946
23. Tan M, Baski A et al (2005) Comparison of pioglitazone and gliclazide in sustaining glycemic control over 2 years in patients with type 2 diabetes. Diabetes Care 28:544–550