
Research Article

Kernel-Based Visual Hazard Comparison (kbVHC): a Simulation-Free Diagnostic for Parametric Repeated Time-to-Event Models

Sebastiaan C. Goulooze,¹ Piry A. J. Väitalo,¹ Catherijne A. J. Knibbe,^{1,2} and Elke H. J. Krekels^{1,3}

Received 6 July 2017; accepted 16 October 2017

Abstract. Repeated time-to-event (RTTE) models are the preferred method to characterize the repeated occurrence of clinical events. Commonly used diagnostics for parametric RTTE models require representative simulations, which may be difficult to generate in situations with dose titration or informative dropout. Here, we present a novel simulation-free diagnostic tool for parametric RTTE models; the kernel-based visual hazard comparison (kbVHC). The kbVHC aims to evaluate whether the mean predicted hazard rate of a parametric RTTE model is an adequate approximation of the true hazard rate. Because the true hazard rate cannot be directly observed, the predicted hazard is compared to a non-parametric kernel estimator of the hazard rate. With the degree of smoothing of the kernel estimator being determined by its bandwidth, the local kernel bandwidth is set to the lowest value that results in a bootstrap coefficient of variation (CV) of the hazard rate that is equal to or lower than a user-defined target value (CV_{target}). The kbVHC was evaluated in simulated scenarios with different number of subjects, hazard rates, CV_{target} values, and hazard models (Weibull, Gompertz, and circadian-varying hazard). The kbVHC was able to distinguish between Weibull and Gompertz hazard models, even when the hazard rate was relatively low (<2 events per subject). Additionally, it was more sensitive than the Kaplan-Meier VPC to detect circadian variation of the hazard rate. An additional useful feature of the kernel estimator is that it can be generated prior to model development to explore the shape of the hazard rate function.

KEY WORDS: model diagnostics; non-linear mixed effect models; pharmacodynamics; pharmacometrics; repeated time-to-event models.

INTRODUCTION

Pharmacometric models are increasingly used to characterize the repeated occurrence of clinical events. Examples from literature include models for emetic episodes, postoperative analgesic events, bone events in Gaucher's disease, and transient lower esophageal sphincter relaxation (1–4). Repeated time-to-event (RTTE) modeling is theoretically superior to alternative methods like time-to-event, which only considers the first event, and count modeling, which treats the events as counts within time intervals (4). This is because RTTE modeling can

take events into account at the time of their occurrence, provided that the event time is not interval-censored. This is especially important when the variation of the hazard—e.g., during the day or after drug administration—is of interest (5,6).

Simulation-based diagnostics are the most commonly used diagnostics to evaluate pharmacometric RTTE models (7). The Kaplan-Meier visual predictive check (VPC) evaluates a model by comparing the observed and simulated Kaplan-Meier survival plots for every n th occurrence of an event (1,4). Another example is a VPC as proposed by Plan *et al.* in which observed and simulated events are discretized as counts within small time intervals (4). Recently, a hazard-based VPC—using non-parametric estimators of the hazard rate—has been proposed as a diagnostic of parametric time-to-event models by Huh and Hutmacher (8). The use of the hazard instead of survival in this diagnostic might allow for a more direct evaluation of the hazard model. However, the hazard-based VPC is also based on simulations.

A disadvantage of the use of simulation-based diagnostics is that one must be able to generate simulations that are representative of the study in which the original data were collected. If data are obtained from a study that includes

Electronic supplementary material The online version of this article (<https://doi.org/10.1208/s12248-017-0162-9>) contains supplementary material, which is available to authorized users.

¹Division of Systems Biomedicine and Pharmacology, Leiden Academic Centre for Drug Research, Leiden University, Einsteinweg 55, 2333, CC, Leiden, The Netherlands.

²Department of Clinical Pharmacy, St. Antonius Hospital, Nieuwegein, The Netherlands.

³To whom correspondence should be addressed. (e-mail: e.krekels@lacdr.leidenuniv.nl)

features such as dose titration or informative dropout, the results from the simulations could be misleading if these features are not correctly accounted for (9). Although this can occasionally be remedied by including additional features in the simulations, this will considerably increase the complexity of the simulation-based analysis. More importantly, when specific study features cannot be correctly accounted for in simulations, the modeler has little options to evaluate a RTTE model. Although residual-based diagnostics have been proposed, their use in the pharmacometric literature is limited (10). Most proposed residuals are only defined for observed events, and their interpretation can be complicated. For example, the martingale and (modified) Cox-Snell residuals can have highly skewed distributions even if the correct model is fitted to a dataset (2). As such, these residuals provide little information on the accuracy of the estimated shape of an underlying hazard function over time (2,10,11).

In this study, we present the kernel-based visual hazard comparison (kbVHC), a novel simulation-free diagnostic for RTTE models. The sensitivity and specificity of the kbVHC to detect model misspecifications were evaluated in simulated scenarios and compared with that of the Kaplan-Meier VPC. Based on these simulations, guidance for the use of the

kbVHC is provided. We also discuss η -shrinkage in the context of RTTE modeling and its influence on the kbVHC.

METHODS

Kernel-Based Visual Hazard Comparison

The kbVHC is designed to evaluate whether the mean predicted hazard rate of a parametric RTTE model is an adequate approximation of the true hazard rate over time. Because the true hazard rate cannot be directly inferred from the observed event data, a non-parametric kernel estimator of the hazard rate is used instead (see below, under “Kernel hazard rate estimator”). The mean hazard rate of the parametric model over time is calculated from the individual posthoc hazard estimates of all non-censored subjects (HAZ_{posthoc}). In the visual output of the diagnostic (see Fig. 1), the HAZ_{posthoc} (solid red line) is plotted *versus* time together with the kernel hazard rate estimate (dashed black line) and its 95% bootstrap confidence interval (gray shaded area). These kbVHC plots can then be inspected for deviations of HAZ_{posthoc} from the non-parametric kernel hazard rate and its 95% bootstrap confidence interval. The

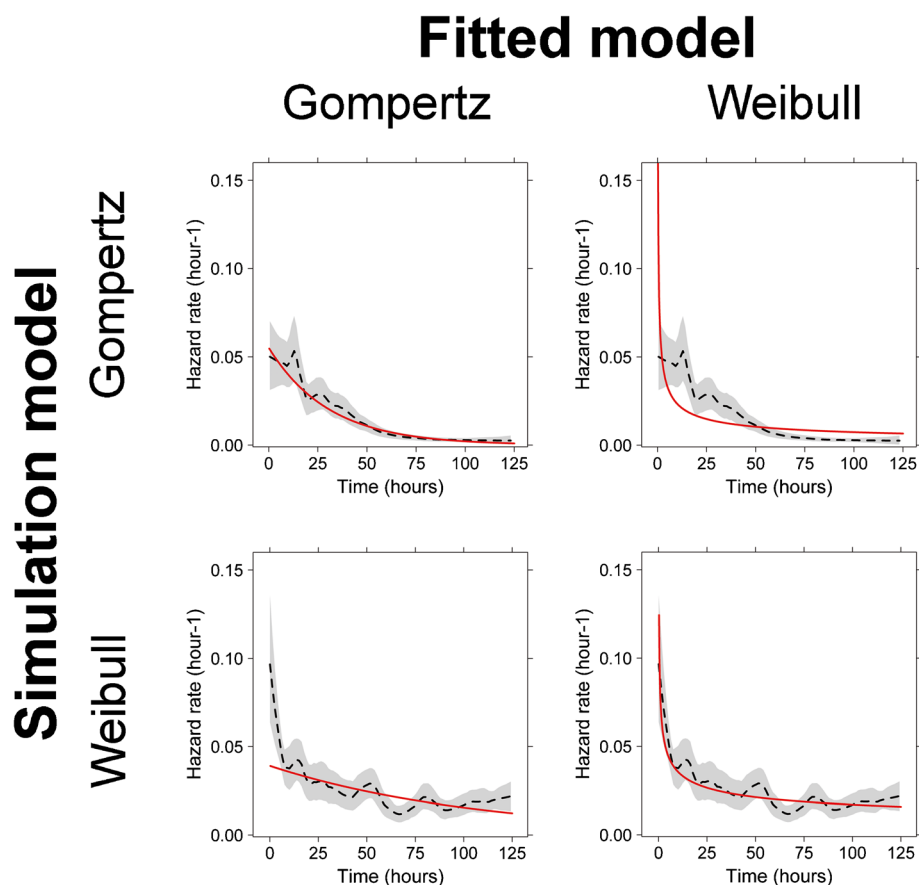


Fig. 1. kbVHC of two simulated datasets fitted with both true and misspecified models. The solid red line shows the HAZ_{posthoc} of the parametric model, the dashed black line with shaded area shows the non-parametric kernel estimate of the hazard rate and its 95% confidence interval, respectively. The two datasets were simulated with 125 subjects and frailty $\omega^2 = 0.75$: one with a Gompertz hazard model (222 events), the other with a Weibull hazard model (379 events). CV_{target} was set at 20%

Kernel-Based Visual Hazard Comparison

method is described below and the developed R code can be found in the Supplemental Material 1.

Kernel Hazard Rate Estimator

The non-parametric kernel estimator of the hazard rate that is used is based on previous work by Chiang *et al.* and Muller and Wang (12,13). The estimator uses the Epanechnikov kernel, with the associated boundary kernel being used to estimate the hazard within one bandwidth of the study start (T_{start}) or study stop time (T_{stop}). The “degree of smoothing” of the kernel hazard rate is determined by the local kernel bandwidth. A smaller bandwidth results in less smoothing of the hazard rate, and vice versa. However, decreasing the bandwidth will generally also increase the uncertainty (expressed as coefficient of variation or CV) of the kernel hazard rate. Here, we set the local kernel bandwidth to the minimal kernel bandwidth that results in CV of the kernel hazard rate that is at or below a user-defined target CV (CV_{target}). This is done by the following steps:

Step 1: Define the search space for the bandwidths.

The user selects the number of time points (N_{time}) at which the minimum satisfactory bandwidth will be determined, and the number of bandwidths (N_{Bw}) that will be tested at these time points.

The N_{time} time points are selected with equidistant distribution between T_{start} and T_{stop} of the study. The tested N_{Bw}

bandwidths are also spread equidistantly between a minimum (Bw_{min}) and maximum (Bw_{max}) bandwidth. Analogue to Muller and Wang (12), Bw_{min} and Bw_{max} are determined by the study duration and the number of observed events in the dataset (N_{obs})

$$Bw_{\text{min}} = \frac{0.2 \times (T_{\text{stop}} - T_{\text{start}})}{8 \times N_{\text{obs}}^{0.2}} \quad (1)$$

$$Bw_{\text{max}} = \frac{10 \times (T_{\text{stop}} - T_{\text{start}})}{8 \times N_{\text{obs}}^{0.2}} \quad (2)$$

Step 2: Determine local kernel bandwidth on each of the N_{time} time points.

A 1000-sample bootstrap of the data is performed to calculate the coefficient of variation of the kernel hazard estimate ($CV_{\text{bootstrap}}$) for all possible combinations of the N_{Bw} test bandwidths and the N_{time} time points. For each individual time point, the minimum bandwidth that satisfies $CV_{\text{bootstrap}} \leq CV_{\text{target}}$ is selected. If none of the N_{Bw} test bandwidths meet this criterion, Bw_{max} is selected for that time point.

Step 3: Determine local kernel bandwidth for all time points.

The minimum satisfactory bandwidths derived at step 2 are smoothed with an Epanechnikov kernel with associated boundary kernels, as described by Muller and Wang (13).

Table I. Overview of Simulated Scenarios for Which kbVHC Was Evaluated

Simulation hazard model	Tested values
All models	$IHAZ(t) = PHAZ(t) \times e^{\eta_i}$
Number of subjects	50, 100, 125, 500
Frailty (ω^2)	0.09, 0.25, 0.5, 0.75, 1.0
Constant hazard (with or without informative dropout)	$PHAZ(t) = \lambda$
	$IDROP = PDROP \times e^{\kappa \times \eta_i}$
λ [h^{-1}]	0.01, 0.05
$PDROP$ [h^{-1}]	0, 0.015
Dropout-free period [h]	36
κ	-2, -1, 0, 1, 2
Gompertz hazard	$PHAZ(t) = \lambda \times e^{\gamma \times t}$
λ [h^{-1}]	0.005, 0.01, 0.025, 0.05
γ [h^{-1}]	-0.005 · ln 2, -0.01 · ln 2, -0.02 · ln 2, -0.05 · ln 2
Weibull hazard	$PHAZ(t) = \lambda \gamma (\lambda t)^{\gamma - 1}$
λ [h^{-1}]	0.01, 0.025, 0.05, 0.1
γ	0.5, 0.7, 0.9
Circadian-varying hazard	$PHAZ(t) = \lambda \times \left[1 + \text{amp} \times \sin\left(\frac{2\pi}{\text{period}} \times (t + \text{phase})\right) \right]$
λ [h^{-1}]	0.005, 0.01, 0.05
amp	0.1, 0.25, 0.5, 0.75
period [h^{-1}]	24
phase [h^{-1}]	0

$IHAZ$ individual hazard, $PHAZ$, population hazard, t time in hours, e^{η_i} frailty term of the i th subject, ω^2 variance of log frailty, λ scale parameter, $IDROP$ individual dropout hazard, $PDROP$ population dropout hazard, κ parameter that controls the relation between frailty of the event of interest (η_i) and the frailty of dropping out, γ shape parameter of the indicated models, amp relative amplitude of circadian variation of hazard rate

This allows the determination of local kernel bandwidths at every time point between T_{start} and T_{stop} . The bandwidth used to smooth the minimum satisfactory bandwidths from step 2 is constant at: $2 \times (T_{\text{stop}} - T_{\text{start}}) / N_{\text{time}}$.

After determining the local kernel bandwidth, this bandwidth is used to determine the non-parametric hazard rate over time. Another 1000-sample bootstrap using the same kernel bandwidths is used to calculate the 95% confidence interval of this hazard rate. The kernel-based non-parametric hazard rate and its 95% confidence interval at each N_{time} are then plotted *versus* time and compared to the mean hazard rate of the parametric model ($\text{HAZ}_{\text{posthoc}}$) over time.

Evaluation of the kbVHC

To evaluate whether the kbVHC would be capable of distinguishing between true and misspecified models, evaluations were performed by simulating RTTE datasets, refitting these datasets with true and misspecified models in NONMEM, and then performing a kbVHC based on the NONMEM output.

RTTE datasets with 1000 subjects and follow-up time of 125 h were simulated in NONMEM version 7.3 (14), using the MTIME method proposed by Nyberg *et al.* to limit computational time (15). The hazard models used in the simulation were

Gompertz, Weibull, circadian-varying hazard, and constant hazard. Log normally distributed frailty (i.e., inter-individual variability of the hazard rate) was included. From these original datasets, subsets were derived by randomly subsampling 500, 125, or 50 subjects without replacement. For the datasets with constant hazard, (informative) dropout scenarios were also generated. Informative dropout was simulated by incorporating a positive or negative association between the frailty to dropout and the frailty for the event of interest. The dropout hazard rate was constant after an initial dropout-free period of 36 h. An overview of the simulated scenarios is provided in Table I.

The simulated datasets were refitted in NONMEM 7.3 with different hazard models. The stochastic approximation expectation-maximization (SAEM) estimation method was used. The objective function value was obtained by performing the expectation step of the importance sampling (IMP) method using the final parameter estimates of the SAEM output (5). All Gompertz and Weibull datasets were fitted with both Weibull and Gompertz hazard models, while datasets with circadian-varying hazard were fitted with either a constant hazard or a circadian-varying hazard model.

After refitting, the kbVHC was constructed as described above using $\text{CV}_{\text{target}}$ values from 5 to 40%. This was done to determine the impact of $\text{CV}_{\text{target}}$ on the degree of smoothing

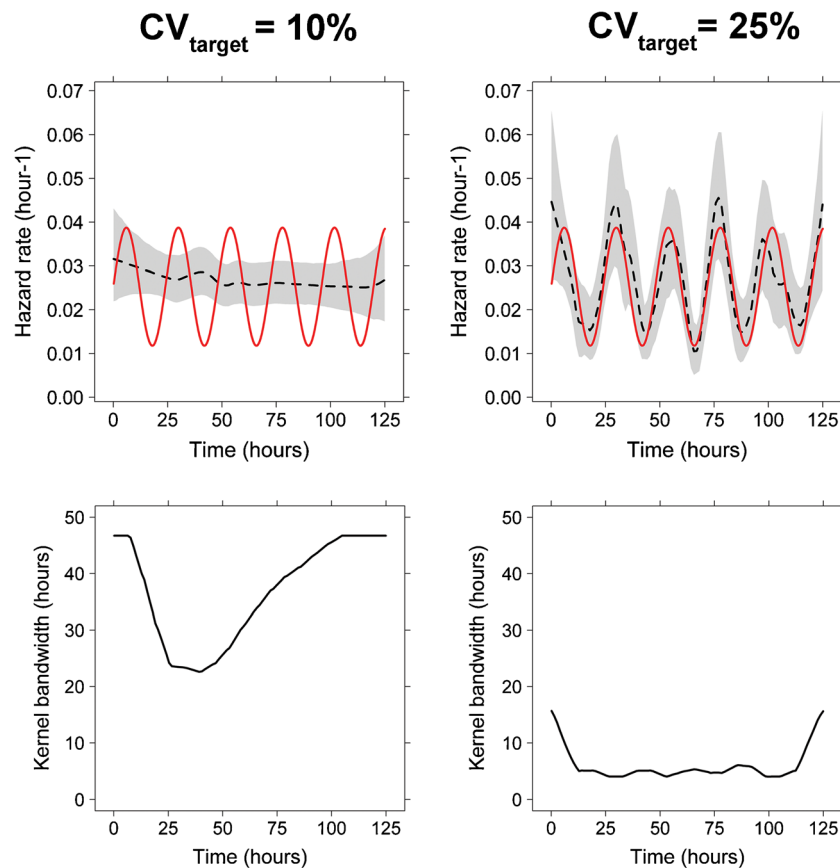


Fig. 2. Impact of user-defined $\text{CV}_{\text{target}}$ on the kbVHC diagnostic in a simulated dataset with 125 subjects and a circadian-varying hazard (415 events). The solid red line in the upper plots shows the $\text{HAZ}_{\text{posthoc}}$; dashed black line with shaded area shows the kernel estimated hazard rate and its 95% confidence interval, respectively. The dataset was refitted with the true circadian-varying hazard model, after which the kbVHC was generated with different values of $\text{CV}_{\text{target}}$. The lower plots show the local bandwidth that was used for the kernel estimator of the hazard

Kernel-Based Visual Hazard Comparison

of the kernel hazard rate and the performance of the kbVHC. N_{time} and N_{BW} were both set to 20, based on previous work (13). For selected illustrative datasets, the performance of the kbVHC was compared to that of the Kaplan-Meier VPC as described by Juul *et al.* (16).

Effect of η -Shrinkage on $\text{HAZ}_{\text{posthoc}}$

The kbVHC makes use of individual post hoc estimates of the hazard for all non-censored subjects to calculate the mean model-predicted $\text{HAZ}_{\text{posthoc}}$. As such, $\text{HAZ}_{\text{posthoc}}$ might be affected by shrinkage of the individual estimate of the frailty term (η -shrinkage). Shrinkage occurs when the data are relatively uninformative on a subject level (e.g., relatively large number of subjects have relatively low number of events). As a result, the individual hazard estimates will shrink towards the population estimate.

In scenarios with large variance of the log normally distributed frailty (ω^2), shrinkage can result in a bias of $\text{HAZ}_{\text{posthoc}}$. However, this bias can be approximated when assuming the individual frailty terms to be log normally distributed, and independent of the typical hazard. This was the case for all scenarios simulated in this study, except the informative dropout scenarios. The approximated bias can

then be used to correct $\text{HAZ}_{\text{posthoc}}$ for the shrinkage-induced bias.

$$\text{BiasHAZ}_{\text{posthoc}} = \frac{\text{mean}(e^\eta) - e^{\frac{\omega^2}{2}}}{e^{\frac{\omega^2}{2}}} \times 100\% \quad (3)$$

$$\text{HAZ}_{\text{posthoc,corrected}} = \text{HAZ}_{\text{posthoc}} \times \frac{e^{\frac{\omega^2}{2}}}{\text{mean}(e^\eta)} \quad (4)$$

To establish the occurrence of bias in the derived $\text{HAZ}_{\text{posthoc}}$ induced by η -shrinkage, we calculated the bias for simulation scenarios that resulted in more than 20% η -shrinkage. In situations where this bias exceeded 10%, we evaluated whether the corrected $\text{HAZ}_{\text{posthoc}}$ could restore the performance of the kbVHC. Unless otherwise specified, the kbVHC plots presented in this paper show the uncorrected $\text{HAZ}_{\text{posthoc}}$.

RESULTS

Figure 1 shows the simulation-free RTTE diagnostic for two simulated datasets based on Weibull and Gompertz models of 125 subjects, which were each fitted with a Weibull

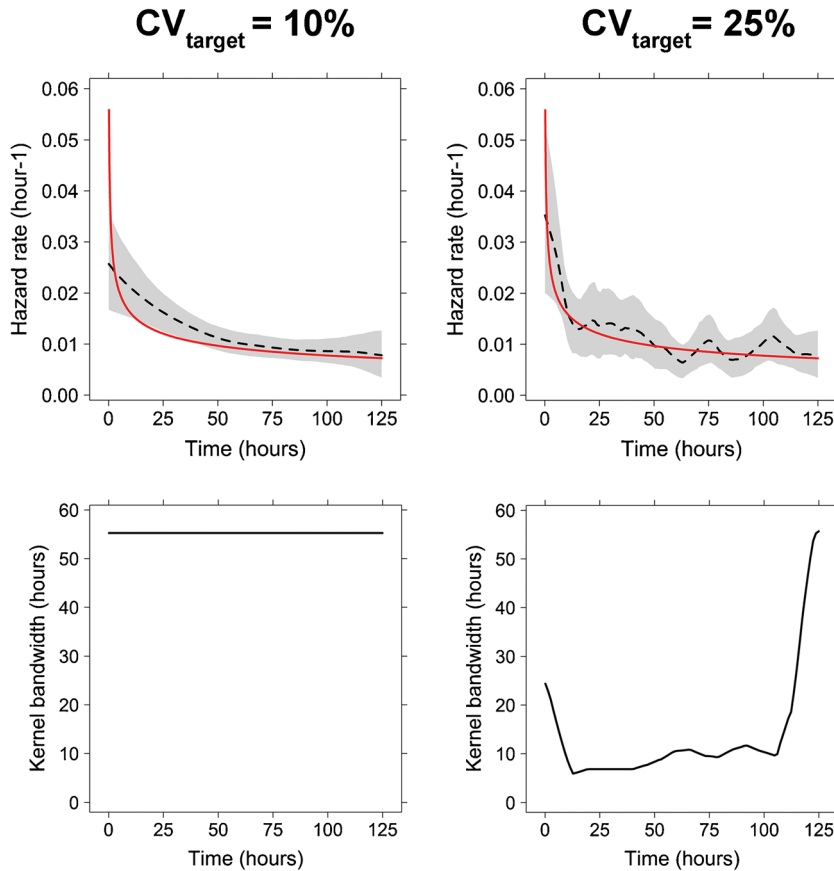


Fig. 3. Comparison of kbVHC with $\text{CV}_{\text{target}}$ of 10 and 25% for a simulated dataset (125 subjects, 179 events) with Weibull hazard. The dataset was refitted with the true Weibull model. The solid red line shows the $\text{HAZ}_{\text{posthoc}}$; dashed black line with shaded area shows the kernel estimated hazard rate and its 95% confidence interval, respectively. The lower plots show the local bandwidth that was used for the kernel estimator of the hazard

and Gompertz model. Fitting the true model to a dataset should result in a model-derived posthoc hazard rate (solid red line) that is comparable to the non-parametric kernel hazard rate (dashed black line). For a misspecified model, we expect to see a deviation from the kernel hazard rate and its confidence interval. It can be seen that HAZ_{posthoc} indeed follows the trend and falls within the 95% confidence interval of the kernel hazard rate when the dataset is fitted with the true model. In case of model misspecification (top right and bottom left), clear deviations of HAZ_{posthoc} from the kernel hazard rate can be observed.

Relation Between CV_{target} and Kernel Bandwidth

To investigate the relationship between CV_{target} , the kernel bandwidth, and the resulting degree of smoothing of

the hazard rate, kbVHCs at various levels of CV_{target} for all simulation scenarios were evaluated. Two examples are presented in Figs. 2 and 3.

Figure 2 shows the kbVHC plots for a dataset with circadian-varying hazard that was fitted with the true model. When CV_{target} is set at 10%, kernel bandwidths are high (>20 h as depicted in the lower plots). As a result, the kernel estimator “smooths over” the circadian variation of the hazard rate. When CV_{target} is increased to 25%, the kernel bandwidths are lowered to around 5 h, and the circadian variation becomes clear in the kernel hazard rate.

Figure 3 shows similar findings for a Weibull dataset fitted with the true Weibull model. When a CV_{target} of 10% is used, the maximum bandwidth (Bw_{max}) is used at every time point (left column, lower plot). As a result, the initial sharp decline of the hazard rate is not visible in the kernel hazard

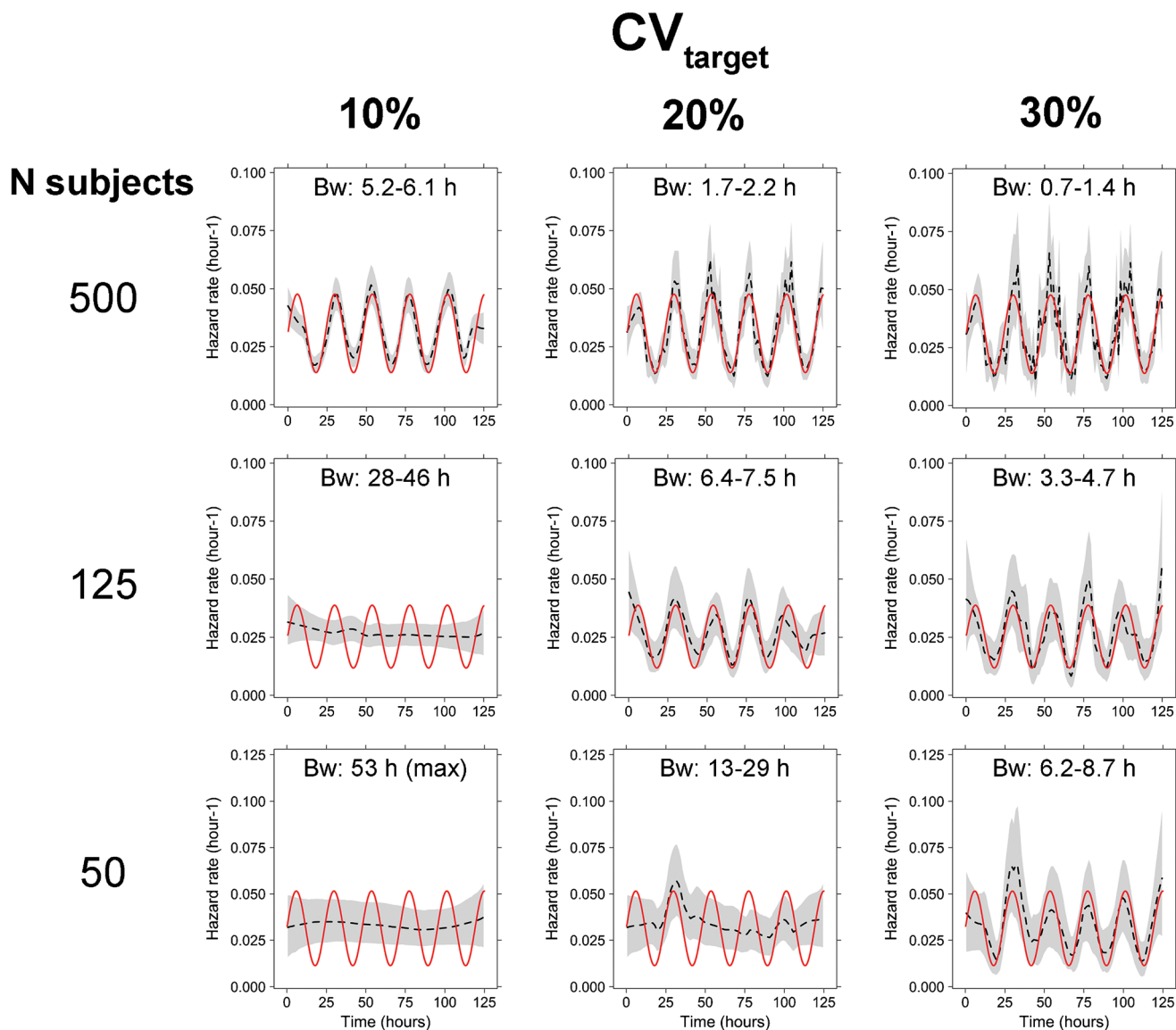


Fig. 4. Degree of smoothing at different values of CV_{target} in circadian-varying hazard datasets with different number of individuals and events. The same circadian-varying hazard model was used to simulate three datasets: **a** 500 subjects, 2007 events; **b** 125 subjects, 415 events; **c** 50 subjects, 206 events. The solid red line shows the HAZ_{posthoc} of the parametric model obtained by refitting the true model; the dashed black line with shaded area shows the non-parametric kernel estimate of the hazard rate and its 95% confidence interval, respectively. Bw = interquartile range of the local kernel bandwidth

Kernel-Based Visual Hazard Comparison

rate. When CV_{target} is increased to 25%, the bandwidth decreases (right column, lower plot) and the HAZ_{posthoc} of the true model matches the kernel hazard rate better.

Influence of the Number of Events on Resulting Bandwidth

Figure 4 shows kbVHC plots for three simulated datasets with 500, 125, and 50 subjects, respectively. All datasets were simulated with the same circadian-varying hazard model and refitted with the true model. The kbVHC was generated using CV_{target} values of 10, 20, or 30%.

When using a CV_{target} of 10%, only the largest dataset (500 subjects, 2007 events) reveals a circadian variation of the hazard rate that matches the model-derived HAZ_{posthoc} . Here, inter-quartile range of the local kernel bandwidths is between 5.2 and 6.1 h. The circadian variation remains clear for this dataset if the CV_{target} is increased to 20 or 30%, but the kernel hazard rate appears to be slightly undersmoothed in these cases. For the dataset with 125 subjects (415 events), the observed kernel hazard rate shows circadian variation when CV_{target} is set to 20 or 30%, but not with a CV_{target} of 10%. With the smallest dataset of 50 subjects (206 events), the CV_{target} needs to be set to at least 30% to reach local bandwidth that reveal circadian variation in the kernel hazard rate. The 24-h circadian rhythm was generally not detected in the kernel hazard rate, when the kernel bandwidth exceeds 10–12 h.

Comparison with Kaplan-Meier VPC

The Kaplan-Meier VPC plots for the misspecified Gompertz and Weibull models from Fig. 1 are shown in Fig. 5. Like the kbVHC, the Kaplan-Meier VPC clearly indicates the model misspecification of fitting a Weibull model to a dataset simulated with a Gompertz model. Most of the observed Kaplan-Meier curve falls outside of the simulated 95% confidence interval of the simulations. However, the misspecification from fitting a Gompertz model to a dataset simulated with a Weibull model is only suggested by the small deviation of the observed Kaplan-Meier curve from the 95% confidence interval in the first 20 h. This misspecification is more pronounced in the kbVHC presented in Fig. 1, where the kernel hazard rate is almost twofold the model-predicted HAZ_{posthoc} in the first few hours.

Kaplan-Meier VPCs were performed for a dataset with 500 subjects and a circadian variation of the hazard with an amplitude of 50% that was refitted with a constant hazard model. As illustrated in Fig. 6, although the observed Kaplan-Meier curve of the first event does appear to oscillate, it remains within the 95% confidence interval of the simulated Kaplan-Meier curve. The figure also shows that the circadian variation of the hazard—and the misspecification of the fitted constant hazard model—is much clearer when the kbVHC is used. For a scenario with a circadian amplitude of 25%, the circadian variation remains visible in the kbVHC, but disappears in the Kaplan-Meier VPC (Supplemental Material 2).

Figure 7 shows a scenario where a constant hazard model is used to fit a dataset simulated with a constant hazard model and informative dropout from 36 h onwards. In this simulated scenario, the hazard of dropping out of the study was negatively

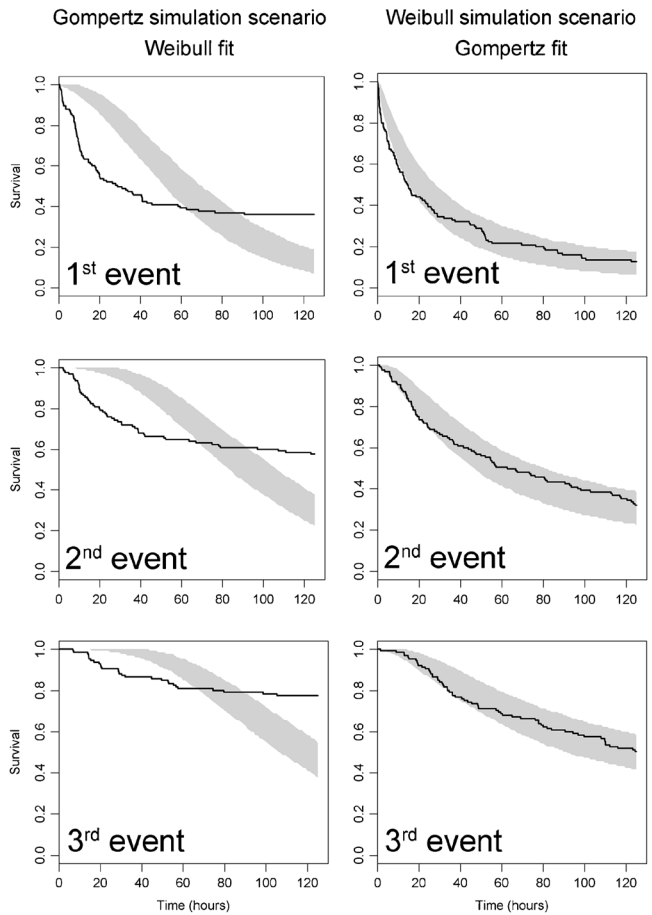


Fig. 5. Kaplan-Meier VPC of misspecified Weibull and Gompertz models. Shown are the observed Kaplan-Meier curves of the original dataset (solid line) and the 95% confidence interval from 1000 simulations from the misspecified model (shaded area). The kbVHC plots for these datasets are shown in Fig. 1

correlated to subject's hazard to experience the event of interest. Without accounting for informative dropout in the simulations, the Kaplan-Meier VPC shows mismatch between most of the observed and simulated Kaplan-Meier curves; at later time points, the simulations overestimate the survival. The kbVHC shows no sign of model misspecification, as both the HAZ_{posthoc} and the kernel hazard estimate increase when the informative dropout starts after 36 h.

Effect of η -Shrinkage on HAZ_{posthoc}

The extent of shrinkage in the parametric model is associated with the amount of events per subject. For example, with an average 1.8 events per subject, the Gompertz fit of the Gompertz scenario shown in Fig. 1 resulted in a shrinkage of 35.6%. In a simulated scenario with a 90% lower typical hazard rate (i.e., average 0.18 events per subject), shrinkage increased to 65.7%. We also observed that shrinkage tends to be asymmetric—subjects with a low hazard were more strongly shrunk towards the population estimate than subjects with a higher than typical hazard. These findings reflect the lower informativeness of the subject's data when that subject has little or no events.

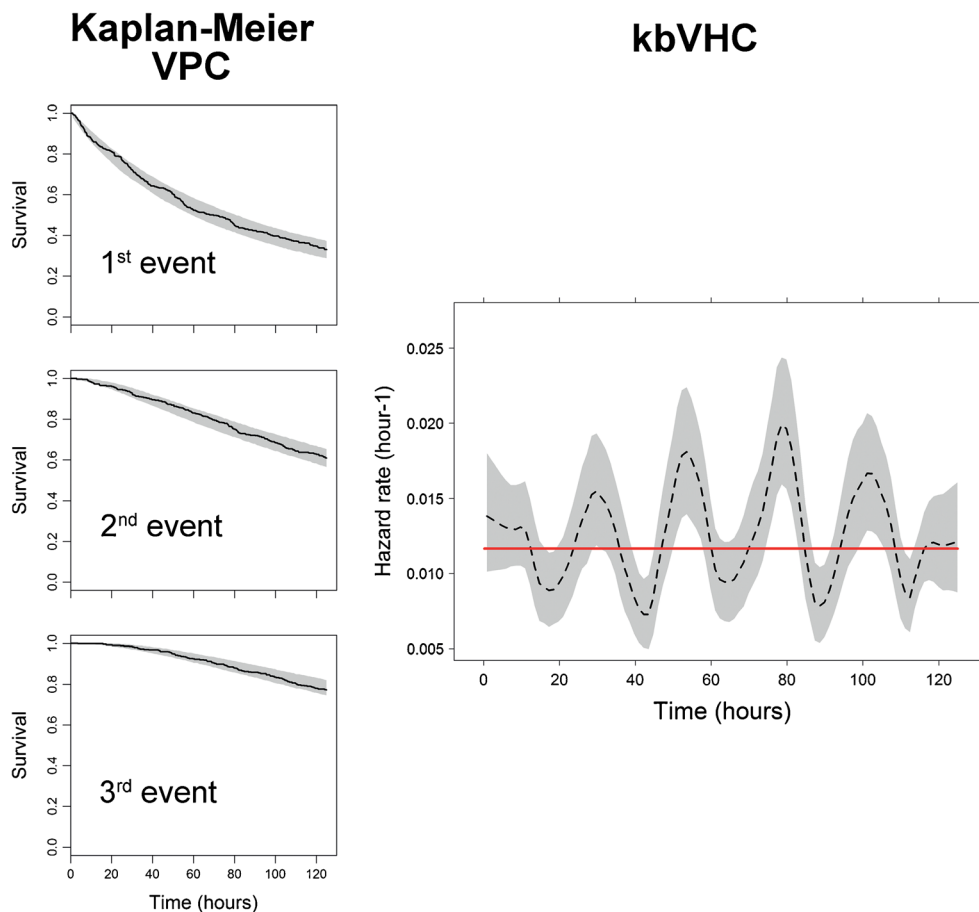


Fig. 6. Comparison of kbVHC and Kaplan-Meier VPC for a scenario with simulated data from a circadian-varying hazard model fitted with a constant hazard model. The simulated dataset contains 500 subjects, 786 events, and 50% amplitude of the circadian variation of the hazard. The Kaplan-Meier VPC plots shows Kaplan-Meier curves for the first, second, and third event within individuals of the original dataset (solid line) and the 95% confidence interval from 1000 simulations from the fitted model (shaded area). The kbVHC plot shows the $\text{HAZ}_{\text{posthoc}}$ of the fitted model (solid red line), and the non-parametric kernel estimate of the hazard rate and its 95% confidence interval (dashed black line with shaded area, respectively). $\text{CV}_{\text{target}} = 15\%$. $\omega^2 = 0.5$

Going back to the previous example, the approximated $\text{HAZ}_{\text{posthoc}}$ bias of the Gompertz fit of the Gompertz scenario in Fig. 1 is only -7.6% , despite relatively high shrinkage (35.6%). Here, $\text{HAZ}_{\text{posthoc}}$ probably does not need to be corrected for an appropriate interpretation of the kbVHC. However, for the scenarios with 65.7% shrinkage, the approximated bias is -26.2% , respectively. As expected, the uncorrected $\text{HAZ}_{\text{posthoc}}$ is generally lower than the kernel estimate of the hazard rate (Fig. 8). Applying the bias correction from Eq. 4 restores agreement between $\text{HAZ}_{\text{posthoc}}$ and the kernel hazard estimate.

DISCUSSION

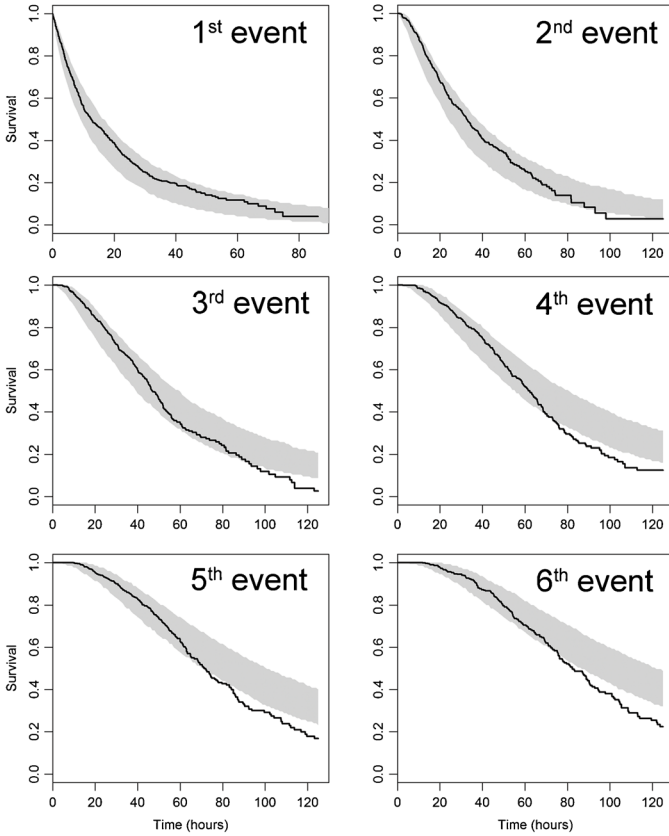
In this study, we developed and evaluated the kbVHC, a simulation-free diagnostic to evaluate the structural submodel of RTTE models in a non-linear mixed-effect setting. The kbVHC can be used to identify misspecification of the model-predicted hazard over time.

Like the hazard-based VPC by Huh and Hutmacher, the kbVHC uses a non-parametric estimate of the hazard rate

(8). However, there are several differences between the two methods. The hazard-based VPC is a diagnostic for time-to-event models, where the non-parametric hazard rate of the observed data is compared to the non-parametric hazard rate of many simulated datasets (8). The kbVHC is primarily intended as a diagnostic for RTTE models that relies on bootstrapping of the observed data to obtain a confidence interval of the non-parametric hazard rate, which is then compared to the mean posthoc hazard of an RTTE model.

In the studies we performed, we showed that the kbVHC was able to identify the correct model in datasets simulated with Gompertz and Weibull models, despite the fact that these models resemble each other in functional form (8). The kbVHC performs comparably to the Kaplan-Meier VPC in some situations while at the same time being easier to interpret (Figs. 1 and 5). In simulated scenarios with a rapidly changing hazard rate (Weibull and circadian-varying hazard), the kbVHC appeared to be more sensitive to model misspecification than the Kaplan-Meier VPC (Figs. 1, 5, and 6).

An additional advantage of the kbVHC over the Kaplan-Meier VPC is that it does not require simulations. These

Kaplan-Meier
VPC

kbVHC

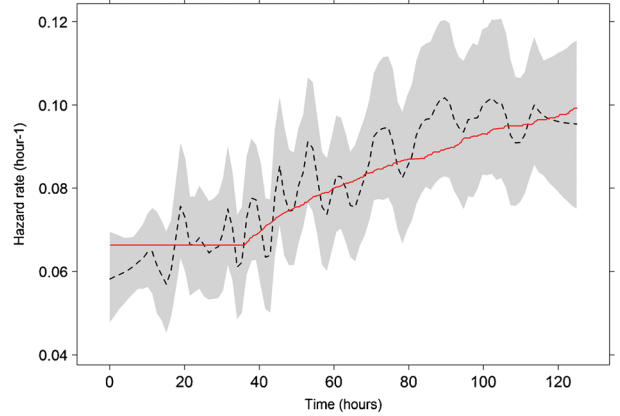


Fig. 7. Comparison of a kbVHC and a Kaplan-Meier VPC that does not account for informative dropout for a scenario with constant hazard with informative dropout ($\kappa = -2$) fitted with a constant hazard model. The dataset contains 500 subjects (at the start of the study) and 3198 events. The Kaplan-Meier VPC plots show Kaplan-Meier curves of the original dataset (solid line) and the 95% confidence interval from 1000 simulations from the fitted model (shaded area). The kbVHC plot shows the $\text{HAZ}_{\text{posthoc}}$ of the parametric model (solid red line), and the non-parametric kernel estimate of the hazard rate and its 95% confidence interval (dashed line with shaded area, respectively). $\text{CV}_{\text{target}} = 10\%$. $\omega^2 = 0.5$

simulations can be difficult and time-consuming to generate, especially in situations with flexible study designs including dose titration or informative dropout (9). When not accounting for these features correctly, the Kaplan-Meier VPC can suggest model misspecification when there is none, as we have shown in Fig. 7 for a scenario with informative dropout. For the same dataset, the kbVHC correctly indicated that there was no model misspecification. It has to be mentioned however that in scenarios without a dropout-free period, there may not be sufficient information to estimate the $\text{HAZ}_{\text{posthoc}}$ of the subjects that dropout early, causing strong asymmetric shrinkage that could introduce bias in the kbVHC.

The computational time needed to generate the kbVHC is in the range of several seconds to minutes, which is considerably faster than the computational time needed to generate the Kaplan-Meier VPC. For example, the bottom-left kbVHC plot in Fig. 1 was generated within a minute, while the simulations for the Kaplan-Meier VPC of this dataset (Fig. 5, right column) took almost an hour (without parallelization). Additionally, calculation of the non-parametric hazard rate and its confidence interval is most

time-consuming, but this needs to be calculated only once for each dataset, as it is independent of the parametric model. An additional advantage of the kernel estimator of the hazard rate being independent of the parametric model is that it can be generated during exploratory data analysis, to inform model development at the start of an analysis. For example, the shape of the kernel hazard rate could inform which parametric hazard models might be appropriate and help determine plausible initial estimates of their parameters.

The degree of smoothing of the kernel estimator depends on a user-defined $\text{CV}_{\text{target}}$, but also on the number of events in the dataset. When $\text{CV}_{\text{target}}$ is too low, the kernel oversmooths the hazard rate, and interesting features of the underlying data (such as circadian rhythm) might be obscured. When the $\text{CV}_{\text{target}}$ is too high, the data will be undersmoothed which would introduce spurious patterns in the kernel-estimated hazard rate. Based on the scenarios shown in Figs. 2, 3, and 4 (and multiple scenarios not shown here), we find the following settings to work well in practice: for datasets with less than 250 events, a suitable range of $\text{CV}_{\text{target}}$ can be 15–40%; for datasets with 250–1000 events, a $\text{CV}_{\text{target}}$ range of 10–30% may work well; and for datasets

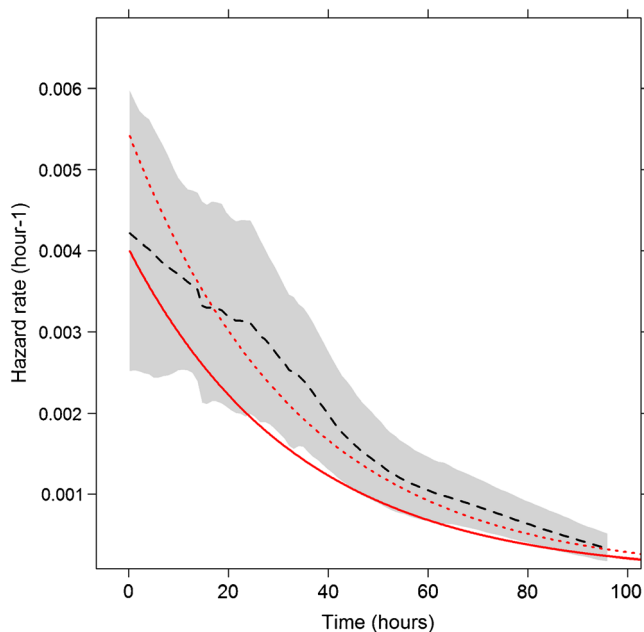


Fig. 8. Comparison of kbVHC using uncorrected (solid red line) or corrected (dotted red line) $HAZ_{posthoc}$ in a scenario with high η -shrinkage (65.7%). The dashed line with shaded area shows the non-parametric kernel estimate of the hazard rate and its 95% confidence interval, respectively. Shown is a kbVHC plot of a Gompertz model fit of a Gompertz scenario with 500 subjects and only 89 events in total. $CV_{target} = 20\%$. $\omega^2 = 0.5$

with more than 1000 events, a CV_{target} between 5 and 20% may work. These ranges provide guidance, but suitable CV_{target} values are context-specific and should be established on a case-by-case basis. The goal is to get the smoothest curve that still captures all important features of the data, similar to other non-parametric “smoothers,” such as the Loess regression (17). This can be done by examining the non-parametric hazard estimates at different CV_{target} values, and using expert knowledge to interpret whether patterns that emerge with less smoothing (i.e., higher CV_{target} values) represent important features of the data or spurious patterns.

The addition of the plot with the local kernel bandwidth over time aids in interpreting the kbVHC, as this will give information on time-varying features of the data that may or may not be captured (see Figs. 2 and 3). For example, with large bandwidths (>15 h), the kernel estimator provides limited information about any 24-h circadian variation of the hazard rate. Alternatively, to specifically test for a 24-h circadian variation, one might consider to fix the bandwidth at 6 to 10 h instead of using the CV_{target} -based approach used in this paper.

As no diagnostic test can evaluate all relevant aspects of a model, the kbVHC should be seen as complimentary to other model diagnostics. Contrary to simulation-based diagnostics like the VPC, where confidence intervals represent variability in simulated data, the confidence interval in the kbVHC represents precision of the kernel hazard estimate at the given kernel bandwidth. Therefore, the confidence interval of the kbVHC cannot be used to evaluate the statistical submodel or the covariate model of a non-linear mixed-effects model. Moreover, the width of the 95% confidence interval of the non-parametric kernel hazard rate

is directly dependent on the user-defined CV_{target} . As a result, the $HAZ_{posthoc}$ of the true model does not necessarily have a 95% probability to be within the 95% confidence interval of the non-parametric hazard rate, as seen in Fig. 2. The confidence interval can nevertheless (qualitatively) aid the user in determining whether patterns or deviations from $HAZ_{posthoc}$ in the non-parametric hazard rate are of spurious or more structural nature.

Like many other model diagnostics that rely on empirical Bayesian estimates, the kbVHC can be affected by high levels of η -shrinkage (9,18). For kbVHC, η -shrinkage can introduce a bias of the model-predicted $HAZ_{posthoc}$. An important advantage of the use of the EBEs is that the kbVHC can in some cases account for the influence of dose titration or informative dropout as demonstrated in Fig. 7. When for instance using the estimated variance of the frailty term to predict the mean hazard rate, this property would be lost. An alternative to using EBEs would be to use random sampling from conditional distribution of the individual parameters, as proposed by Lavielle and Ribba (19); however, these are not readily available in NONMEM 7.3 output.

In most tested scenarios, this bias was low ($<10\%$) even when there were moderate levels of η -shrinkage ($\pm 30\%$). Higher bias was observed for scenarios with a low average number (<1) of events per subject and high variance of the frailty term ($\omega^2 \geq 0.5$). To maintain acceptable performance of the kbVHC in these situations, we have proposed a simple correction method for log normally distributed frailty terms. Since frailty is often log normally distributed in pharmacometric RTTE models, we anticipate that this correction method will increase the usefulness of the kbVHC in practice. It should however be mentioned that in situations with informative dropout or dose titration, the frailty terms are not independent of the typical hazard, and the proposed correction method should not be used.

Additionally, pharmacometric RTTE models will often be linked to a pharmacokinetic model, with the predicted drug concentration affecting the model-predicted hazard rate. EBE shrinkage in the pharmacokinetic parameters can affect the predicted drug concentrations, thereby introducing bias in $HAZ_{posthoc}$ and the resulting kbVHC output. The same limitation is to be expected in simulation-based RTTE model diagnostics from a sequentially performed pharmacodynamics analysis in which pharmacokinetic parameters are fixed to their individual (EBE-based) values. With simulation-based RTTE diagnostics, this limitation can be negated by also including inter-individual variability in the pharmacokinetic parameters for the simulations, but for the kbVHC, this is not possible.

As is the case for all pharmacometric visual diagnostics, the evaluation of the kbVHC is mostly qualitative. This limited the feasibility of analyzing and reporting on many repetitions of similar scenarios. We did, however, evaluate the kbVHC in many scenarios (see Table I), and a selection of the most illustrative plots was made for this paper.

CONCLUSION

We have developed a simulation-free diagnostic for RTTE models based on a non-parametric kernel estimation of the hazard rate derived from observed events. The kbVHC

Kernel-Based Visual Hazard Comparison

has a good sensitivity for structural model misspecification, even outperforming the existing Kaplan-Meier VPC for time-varying hazard models. Because the kbVHC does not require simulations, it can also be used in situations where appropriate simulations are difficult to generate. Like other diagnostics that rely on empirical Bayesian estimates, the kbVHC can be affected by high levels of η -shrinkage through a biased HAZ_{posthoc}. However, we found that this bias can be approximated and corrected for, when the model includes log normally distributed frailty. An additional useful feature of the kernel estimator is that it can already be generated prior to model development to explore the shape of the hazard rate function. These advantages make the kbVHC a valuable addition to the diagnostic toolbox for RTTE models.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

REFERENCES

1. Juul RV, Rasmussen S, Kreilgaard M, Christrup LL, Simonsson US, Lund TM. Repeated time-to-event analysis of consecutive analgesic events in postoperative pain. *Anesthesiology*. 2015;123(6):1411–9. <https://doi.org/10.1097/ALN.0000000000000917>.
2. Cox EH, Veyrat-Follet C, Beal SL, Fuseau E, Kenkare S, Sheiner LBA. Population pharmacokinetic-pharmacodynamic analysis of repeated measures time-to-event pharmacodynamic responses: the antiemetic effect of ondansetron. *J Pharmacokinet Biopharm*. 1999;27(6):625–44. <https://doi.org/10.1023/A:1020930626404>.
3. Vigan M, Stirnemann J, Mentre F. Evaluation of estimation methods and power of tests of discrete covariates in repeated time-to-event parametric models: application to Gaucher patients treated by imiglucerase. *AAPS J*. 2014;16(3):415–23. <https://doi.org/10.1208/s12248-014-9575-x>.
4. Plan EL, Ma G, Nagard M, Jensen J, Karlsson MO. Transient lower esophageal sphincter relaxation pharmacokinetic-pharmacodynamic modeling: count model and repeated time-to-event model. *J Pharmacol Exp Ther*. 2011;339(3):878–85. <https://doi.org/10.1124/jpet.111.181636>.
5. Karlsson KE, Plan EL, Karlsson MO. Performance of three estimation methods in repeated time-to-event modeling. *AAPS J*. 2011;13(1):83–91. <https://doi.org/10.1208/s12248-010-9248-3>.
6. Plan EL. Modeling and simulation of count data. *CPT Pharmacometrics Syst Pharmacol*. 2014;3(8):e129. <https://doi.org/10.1038/psp.2014.27>.
7. Plan EL. Pharmacometric methods and novel models for discrete data. Uppsala University 2011. Available from: <http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-150929>. Accessed 8 Aug 2016.
8. Huh Y, Hutmacher MM. Application of a hazard-based visual predictive check to evaluate parametric hazard models. *J Pharmacokinet Pharmacodyn*. 2016;43(1):57–71. <https://doi.org/10.1007/s10928-015-9454-9>.
9. Karlsson MO, Savic RM. Diagnosing model diagnostics. *Clin Pharmacol Ther*. 2007;82(1):17–20. <https://doi.org/10.1038/sj.clpt.6100241>.
10. Holford N. A time to event tutorial for pharmacometricians. *CPT Pharmacometrics Syst Pharmacol*. 2013;2(5):e43. <https://doi.org/10.1038/psp.2013.18>.
11. Collett D. Model checking in the Cox regression model. In: Collett D, editor. *Modelling survival data in medical research*. Boca Raton, FL: CRC Press; 2015. p. 131–70.
12. Chiang CT, Wang MC, Huang CY. Kernel estimation of rate function for recurrent event data. *Scand Stat Theory Appl*. 2005;32(1):77–91. <https://doi.org/10.1111/j.1467-9469.2005.00416.x>.
13. Muller HG, Wang JL. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*. 1994;50(1):61–76. <https://doi.org/10.2307/2533197>.
14. Beal SL, Sheiner LB, Boeckmann AJ, Bauer RJ. *NONMEM Users Guides*. Ellicott City: Icon Development Solutions; 2010.
15. Nyberg J, Karlsson KE, Jönsson S, Simonsson USH, Karlsson MO, Hooker AC. Simulating large time-to-event trials in NONMEM. Population Approach Group Europe. 2014. Available from: <https://www.page-meeting.org/default.asp?abstract=3166>. Accessed 20 May 2016.
16. Juul RV, Nyberg J, Lund TM, Rasmussen S, Kreilgaard M, Christrup LL, *et al*. A pharmacokinetic-pharmacodynamic model of morphine exposure and subsequent morphine consumption in postoperative pain. *Pharm Res*. 2016;33(5):1093–103. <https://doi.org/10.1007/s11095-015-1853-5>.
17. Jacoby WG. Loess: a nonparametric, graphical tool for depicting relationships between variables. *Elect Stud*. 2000;19(4):577–613.
18. Savic RM, Karlsson MO. Importance of shrinkage in empirical bayes estimates for diagnostics: problems and solutions. *AAPS J*. 2009;11(3):558–69. <https://doi.org/10.1208/s12248-009-9133-0>.
19. Lavielle M, Ribba B. Enhanced method for diagnosing pharmacometric models: random sampling from condition distributions. *Pharm Res*. 2016;33(12):2979–88. <https://doi.org/10.1007/s11095-016-2020-3>.