CrossMark

Research Article

Covariates in Pharmacometric Repeated Time-to-Event Models: Old and New (Pre)Selection Tools

Sebastiaan C. Goulooze,[1] Elke H. J. Krekels,[1] Thomas Hankemeier,[1] and Catherijne A. J. Knibbe[1,2,3]

Abstract. During covariate modeling in pharmacometrics, computational time can be reduced by using a fast preselection tool to identify a subset of promising covariates that are to be tested with the more computationally demanding likelihood ratio test (LRT), which is considered to be the standard for covariate selection. There is however a lack of knowledge on best practices for covariate (pre)selection in pharmacometric repeated time-to-event (RTTE) models. Therefore, we aimed to systematically evaluate the performance of three covariate (pre)selection tools for RTTE models: the likelihood ratio test (LRT), the empirical Bayes estimates (EBE) test, and a novel Schoenfeld-like residual test. This was done in simulated datasets with and without a "true" time-constant covariate, and both in the presence and absence of high EBE shrinkage. In scenarios with a "true" covariate effect, all tools had comparable power to detect this effect. In scenarios without a "true" covariate effect, the false positive rates of the LRT and the Schoenfeld-like residual test were slightly inflated to 5.7% and 7.2% respectively, while the EBE test had no inflated false positive rate. The presence of high EBE shrinkage (> 40%) did not affect the performance of any of the covariate (pre)selection tools. We found the EBE test to be a fast and accurate tool for covariate preselection in RTTE models. The novel Schoenfeld-like residual test proposed here had a similar performance in the tested scenarios and might be applied more readily to time-varying covariates, such as drug concentration and dynamic biomarkers.

KEY WORDS: covariate model building; empirical Bayes estimate; non-linear mixed effects modeling; repeated time-to-event.

INTRODUCTION

Repeated time-to-event (RTTE) models characterize the hazard or event rate of clinically relevant events, such as rescue analgesia events, emetic episodes, and bone events in Gaucher's disease (1–3). An important part of RTTE modeling is the search for predictors of this hazard. These predictors, also known as covariates, help us understand and predict differences in the observed events between individuals (4). Despite the importance of covariates, there is limited literature available on how to develop the covariate submodel in RTTE models (2).

[1] Division of Systems Biomedicine and Pharmacology, Leiden Academic Centre for Drug Research, Leiden University, Room number 0.2.11, Einsteinweg 55, 2333 CC, Leiden, The Netherlands.
[2] Department of Clinical Pharmacy, St. Antonius Hospital, Nieuwegein, The Netherlands.
[3] To whom correspondence should be addressed. (e–mail: c.knibbe@antoniusziekenhuis.nl)

Whether or not the inclusion of a particular covariate significantly improves the RTTE model is commonly tested with the likelihood ratio test (LRT), which tests the difference in likelihood of a model with and without inclusion of a covariate relationship for statistical significance (5,6). However, the performance of LRT as a method for covariate selection in RTTE models has only been evaluated for binary covariates (2). Additionally, covariate selection using LRT is time-consuming, as it requires a separate model to be fitted for each tested covariate relationship at each step of the forward inclusion procedure (7). Therefore, it can be helpful to use techniques for the preselection of promising covariates, which are later formally tested for significance using LRT (Fig. 1). By only applying the LRT to the most promising covariates, the overall computational time required for covariate modeling is reduced.

Preselection of covariates in pharmacometric modeling is commonly done by evaluating the association between covariates and the empirical Bayesian estimates (EBE) of subjects' deviations from the population parameters (8). The EBEs are dependent on the level of subject-specific information and shrink towards zero if this information is sparse or absent. It has been suggested that with levels of EBE shrinkage above 20–30%, the EBEs are less suitable for
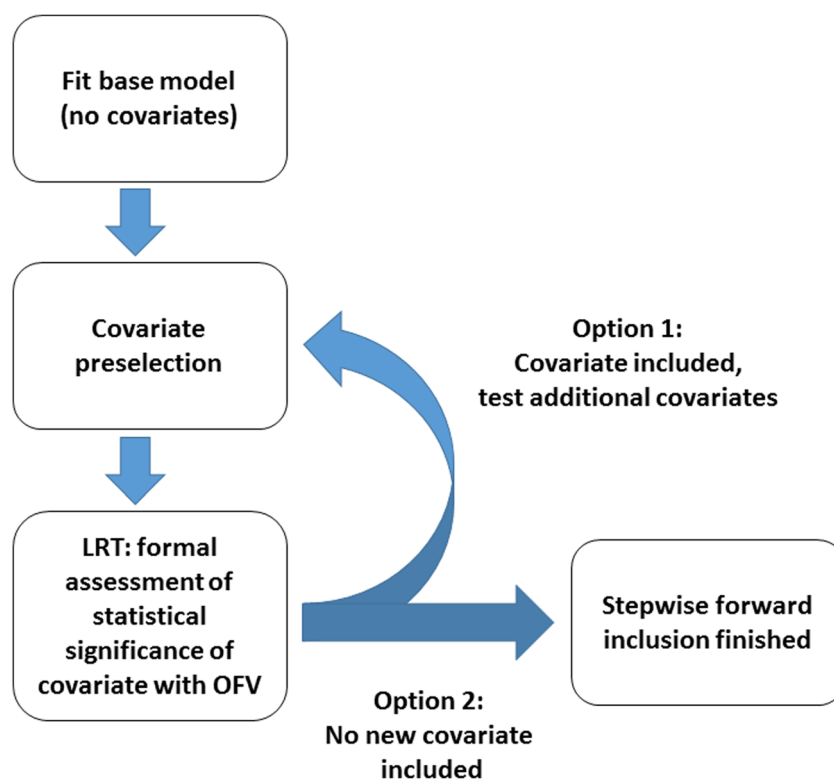
Fig. 1. Schematic representation of how stepwise forward inclusion of covariates with a covariate preselection tool is generally used in practice. The use of a preselection tool is performed to limit the number of covariate effects that are tested for statistical significance using the time-consuming LRT. During the stepwise forward selection, covariates are added to the model one by one, until none of the preselected covariates result in a statistical significant improved model and the stepwise forward inclusion is finished. ΔOFV = difference in objective function value after inclusion of a particular covariate in the model

covariate selection (9), although recent publications have argued against this (5,10). However, the performance of the EBE test and the impact of EBE shrinkage on this test are unknown for RTTE modeling, as this test has yet to be evaluated as a covariate preselection technique for these models.

Considering the potential shrinkage-related issues with the EBE test, we developed a Schoenfeld-like residual test, as a novel covariate preselection technique for pharmacometric RTTE models. The Schoenfeld-like residual is the difference between the observed and expected covariate value of a given event at a given time (11,12). Because it considers covariate values at the time of each event, the residual does not require covariates to be time-constant. The Schoenfeld-like residual used here is adapted from a similar residual proposed by Wileyto *et al.* for parametric time-to-cure models (12).

Here, we aimed to systematically evaluate the performance of three covariate (pre)selection techniques for RTTE models: the likelihood ratio test (LRT), the empirical Bayes estimates (EBE) test, and a novel Schoenfeld-like residual test. The performance of the three techniques was evaluated head-to-head by quantifying the false positive rate (Type I error), and the true positive rate or statistical power in various simulation scenarios. Based on our findings, we provide practical guidance on covariate modeling in RTTE models.

METHODS

RTTE Simulation and Model

In this study, RTTE datasets were simulated using models with or without a true effect of a covariate on the hazard rate. These datasets were used to evaluate whether the (pre)selection techniques are able to correctly identify the presence or absence of a true covariate effect. Table I provides an overview of all simulated scenarios.

In *R*, simulation input datasets were generated with a single binary or continuous covariate. The binary covariate was coded as either a 0 or a 1, with each value occurring with a 50% frequency in the population. The continuous covariate for each individual in the dataset was sampled from a standard normal distribution (mean of 0, standard deviation of unity). RTTE trials with these patient cohorts were then simulated in NONMEM 7.3 using the MTIME method proposed by Nyberg *et al.* (13,14). Follow-up time was kept constant at 125 h, with no dropout. A constant hazard model, also known as exponential survival model in the context of time-to-event modeling, was used to characterize the instantaneous hazard or event rate of each individual subject. The impact of the binary and continuous covariate on the hazard was modeled according to Eqs. 1 and 2, respectively.

Table I. Overview of Simulation Scenarios

Scenario	Covariate type	# of subjects	Effect$_{cov}$	h_{pop} (h^{-1}) [EBE shrinkage level]	# of datasets simulated
1	Binary	150	0	0.05 [low]	500
2	Binary	150	0	0.005 [high]	500
3	Binary	50	0	0.05 [low]	500
4	Binary	50	0	0.005 [high]	500
5	Continuous	150	0	0.05 [low]	500
6	Continuous	150	0	0.005 [high]	500
7	Continuous	50	0	0.05 [low]	500
8	Continuous	50	0	0.005 [high]	500
9	Binary	150	0.25	0.05 [low]	250
10	Binary	150	0.25	0.005 [high]	250
11	Binary	150	0.35	0.05 [low]	250
12	Binary	150	0.35	0.005 [high]	250
13	Binary	150	0.5	0.05 [low]	250
14	Binary	150	0.5	0.005 [high]	250
15	Binary	150	0.8	0.05 [low]	250
16	Binary	150	0.8	0.005 [high]	250
17	Continuous	150	0.1	0.05 [low]	250
18	Continuous	150	0.1	0.005 [high]	250
19	Continuous	150	0.15	0.05 [low]	250
20	Continuous	150	0.15	0.005 [high]	250
21	Continuous	150	0.25	0.05 [low]	250
22	Continuous	150	0.25	0.005 [high]	250
23	Continuous	150	0.5	0.05 [low]	250
24	Continuous	150	0.5	0.005 [high]	250

EBE, empirical Bayesian estimates; Effect$_{cov}$, covariate effect parameter; h_{pop}, value of population hazard

$$HAZ_{TV}(i) = h_{pop} * (1 + cov_i * Effect_{cov}) \quad (1)$$

$$HAZ_{TV}(i) = h_{pop} * e^{(cov_i * Effect_{cov})} \quad (2)$$

where $HAZ_{TV}(i)$ is the typical hazard of the ith subject with the given covariate value, h_{pop} is the time-constant population hazard for the event of interest of a subject with a covariate value of zero, cov_i is the value of the covariate in subject i, and $Effect_{cov}$ is the parameter that represents the covariate effect on the hazard. Both h_{pop} and the value of the covariates did not vary over time in this study.

In pharmacometric RTTE models, the (unexplained) between-subject variability of the hazard rate (or frailty) is typically modeled as a log-normally distributed term describing the deviations of individual subjects from the population hazard, as shown in Eq. 3.

$$h(i) = HAZ_{TV}(i) * e^{\eta i} \quad (3)$$

where $h(i)$ represents the individual hazard of the ith subject and $e^{\eta i}$ represents the empirical Bayesian estimate (EBE) of the hazard of the ith subject relative to HAZ_{TV}. The degree of between-subject variability was set to a coefficient of variation of 75% in all scenarios, which corresponds with a variance of η of 0.45. h_{pop} was set to either 0.05 or 0.005 h^{-1},

to simulate scenarios with either a relatively high or low number of events per subject, respectively. These values are expected to result in a low or high degree of shrinkage of the EBEs, as we have previously shown shrinkage to be associated to the amount of events per subject, likely because a lower amounts of events in an individual yields less information on an individual level [15].

In simulated scenarios without a true covariate effect (Effect$_{cov}$ = 0), the false positive rate was quantified as the percentage of datasets for which a significant association ($p < 0.05$) between the covariate and the hazard is reported by the covariate (pre)selection technique. We generated 500 datasets for each of eight different scenarios without a true covariate effect, in which each scenario had a unique combination of h_{pop} value (0.05 or 0.005 h^{-1}), number of subjects (150 or 50), and type of covariate (binary or continuous). In 16 different simulated scenarios with a true covariate effect, the power of the techniques to detect this covariate effect was quantified. Two hundred fifty datasets with 150 subjects each were generated for each of the 16 scenarios. These 16 scenarios differed from each other by h_{pop} values (0.05 or 0.005 h^{-1}), type of covariate (binary or continuous), and Effect$_{cov}$ (0.25, 0.35, 0.5, and 0.8 for binary covariate; 0.1, 0.15, 0.25, 0.5 for continuous covariates). After the simulation step, all datasets were fitted in NONMEM with a base model that did not include an estimated effect of the covariate (Effect$_{cov}$ fixed to zero). The output of these base model fits was then used as an input for the covariate preselection techniques described below. To perform the LRT, the fit of a model,

where Effect$_{cov}$ was estimated, was compared to the base model fit.

Covariate Selection with LRT

The LRT is performed by comparing the objective function value (-2 log-likelihood) between a model with an estimated covariate effect (Effect$_{cov}$) and a base model without a covariate effect (Effect$_{cov}$ fixed to 0). When the inclusion of the estimated covariate effect results in a drop in objective function value of at least 3.84 points, the covariate was considered to be significant with $p < 0.05$. To test and compare type I error rates and statistical power of all techniques, in this study, the LRT was applied to each of the simulated datasets, regardless of the results of the covariate preselection techniques described below.

Covariate Preselection

Empirical Bayes Estimates Test

All fitted models included an estimated frailty term, which represents the magnitude of between-subject variability around the population hazard estimate. During the post hoc step in NONMEM, EBEs of the frailty term are generated for each individual subject (ηi in Eq. 3). For scenarios with a continuous covariate, the correlation between EBEs from the NONMEM output and continuous covariate values was calculated using the Pearson correlation test. For binary covariates, an unpaired two-way t test was used to test for significant differences between the two groups ($\alpha = 0.05$).

Schoenfeld-Like Residual Test

The Schoenfeld-like residual is a residual that is defined for each separate event and each covariate. It is defined by the difference between the expected and observed covariate value of the given event (Eq. 4). The observed covariate value is the covariate value of the subject experiencing the event at the time of the event. The expected covariate value is defined as the weighted average of the covariate values of all non-censored subjects at the time of the event, where the weight is the population hazard of each subject (Eq. 5).

$$SF_j = cov_j - predicted\_cov_j \qquad (4)$$

$$predicted\_cov_j = \frac{\sum_{i=1}^{k} cov_i * HAZ_{TV}(i)}{\sum_{i=1}^{k} HAZ_{TV}(i)} \qquad (5)$$

where SF_j is the Schoenfeld-like residual for the jth event, COV_j is the covariate value of the subject experiencing the jth event at the respective event time, $predicted\_cov_j$ the predicted covariate value at the time of the jth event, k is the number of uncensored subjects at the time of the jth event, cov_i is the covariate value of the ith uncensored subject, HAZ_{TVi} is the population hazard of the ith uncensored subject.

In the absence of a (true) covariate effect, the expected mean value of the Schoenfeld-like residual is zero (11,16). The 95% confidence interval of the mean Schoenfeld-like residuals was calculated with a 1000 sample bootstrap of the model output dataset; resampling subjects with replacement and then recalculating the mean Schoenfeld-like residual in each resampled dataset. If the 95% confidence interval of the mean Schoenfeld-like residual did not include zero, the covariate was considered to be significantly associated with the hazard rate ($p < 0.05$).

Evaluation of the Covariate (Pre)Selection Techniques

For simulated scenarios without a true covariate effect (Effect$_{cov} = 0$), the false positive rate and its 95% confidence interval were calculated using the *prop.test* function in *R* (one-sample proportion test with continuity correction). This was done for each scenario separately, but also in a pooled assessment where all simulated datasets without a true covariate effect were analyzed together to increase the power to detect inflations of the false positive rate. For simulation scenarios with a covariate effect, the *prop.test* function was used to compare the power of the (pre)selection techniques to detect the covariate effect (two-sample test for equality of proportions with continuity correction with $\alpha = 0.05$). This pairwise comparison of the power was performed between LRT versus EBE test, LRT versus Schoenfeld-like residual test, and EBE test versus Schoenfeld-like residual test. This was done for each scenario separately, but also in a pooled assessment where all simulated datasets with a true covariate effect were analyzed together to increase the power to detect differences in the power. The level of agreement between the two preselection techniques and the LRT was assessed by calculating the percentages of the datasets in which the techniques came to the same conclusion (i.e., significant covariate effect or not). Finally, we compared the computational time needed to run the (pre)selection techniques.

RESULTS

In this work, we evaluated a covariate selection technique (LRT), and two covariate preselection techniques (EBE test and Schoenfeld-like residual test) in 24 different scenarios. The median EBE shrinkage of the base model fits of each simulation scenario was calculated to determine the shrinkage obtained in the simulations with high or low value of the typical hazard h_{pop} (0.05 or 0.005 h^{-1}). Setting the typical hazard h_{pop} to a relatively high value (0.05 h^{-1}) resulted in simulated scenarios with relatively low EBE shrinkage ranging from 11 to 16%. In scenarios with a tenfold lower value of h_{pop} (0.005 h^{-1}), median EBE shrinkage ranged from 42 to 55%.

The false positive rate was evaluated in eight scenarios with 500 datasets of 150 or 50 individuals each, in which there was no true covariate effect (Effect$_{cov} = 0$). In the scenarios with the binary covariate presented in Table II, the Schoenfeld-like residual test has an increased false positive rate in the low shrinkage scenario with 150 subjects (95% CI of 6.2–11.3), but not in the other three scenarios. The LRT and EBE test did not result in a false positive rate significantly different from 5% in any of the separate binary

covariate scenarios. Table III shows the results for the four scenarios in which a continuous covariate was tested. With 50 subject and low shrinkage, all techniques show an inflated false positive rate above 5% ($p < 0.05$). In the three remaining scenarios for the continuous covariate, the false positive rate was not inflated for any of the (pre)selection techniques ($p > 0.05$).

There were no clear trends in false positive rate in either high versus low shrinkage, 150 versus 50 subjects, or binary versus continuous covariate. Only after pooling the results from all eight scenarios (total of 4000 datasets) in Table II and III, was it found that the false positive rates were significantly inflated for both the Schoenfeld-like residual test (7.2%, CI 6.45–8.08), and the LRT (5.7%, CI 5.04–6.51). The false positive rate of the EBE test was not significantly higher than 5% (5.3%, CI 4.60–6.00) in the pooled analysis.

The power to detect a true covariate relationship was quantified in the 16 scenarios in which a covariate effect was included in the simulation. Figure 2 shows the power to detect a true covariate effect of the covariate (pre)selection techniques to increase with increasing covariate values. This power is up to 0.38 lower in the scenarios with high EBE shrinkage compared to similar scenarios with low shrinkage, with no apparent differences in the impact of shrinkage amongst the covariate (pre)selection techniques. The power of the two preselection tools and the LRT to detect the covariate effect were not statistically significantly different in any of the scenarios ($p > 0.05$). Even when the datasets from all 16 scenarios with a "true" covariate effect were pooled ($n = 4000$), no significant difference could be detected between any of the covariate (pre)selection techniques.

Across all 8000 simulated datasets, the preselection techniques were in strong agreement with the LRT on the statistical significance of the covariate effect: 99.2% for the EBE test and 95.4% for the Schoenfeld-like residual test. In the cases where preselection techniques gave a different answer than the LRT, we examined which of the two correctly identified the true model, i.e., with or without a covariate effect. From the 0.8% of cases where the EBE test disagreed with the LRT, the EBE test was correct 50% of the time, with the LRT being correct in the other 50% of the cases. Within the 4.6% disagreement between the Schoenfeld-like residual test and the LRT, the Schoenfeld-like residual test was correct in 38% of the cases, and the LRT in 72% of the cases.

Both preselection techniques, EBE and Schoenfeld-like residual test, required little computational time (within 2 s per dataset) when compared to the LRT (average time of 4.3 min across all scenarios). The EBE test only requires that a correlation test is performed, which can be done directly on the NONMEM output. Although, the generation of the Schoenfeld-like residual test required more code than the EBE test (Supplemental Information), the computation time is greatly reduced compared to the LRT.

DISCUSSION

In this study, we evaluated the performance of three techniques in covariate (pre)selection for pharmacometric RTTE models. In particular, we tested the false positive rate and the power to detect "true" covariate effect in each of these techniques. Two of the evaluated techniques, the EBE test and LRT are commonly used for covariate preselection and selection in other types of pharmacometric models, such as pharmacokinetic models. Additionally, we evaluated a third technique that has hitherto not been used in literature on pharmacometric RTTE models, which uses a Schoenfeld-like residual to preselect promising covariates.

The Schoenfeld-like residual that was used here differs from the originally defined Schoenfeld residual, which is used to test the proportional hazard assumption of Cox regression models. As shown in Eq. 5, the weighted average of the covariate values of the non-censored subjects is obtained using the population hazard of a parametric hazard model with an estimated frailty term. Such a population hazard is not defined in the semi-parametric context in which the original Schoenfeld residual is used. Another difference with the original Schoenfeld residual is the proposed application of the Schoenfeld-like residual: the originally defined Schoenfeld residual is used to test the proportional hazard assumption of included predictors in the model. This is done by visual inspection of the (scaled) Schoenfeld residual over time for any trends. Here, we propose that the Schoenfeld-like residual test can be used to preselect covariates before using the more time-consuming LRT.

Both preselection techniques (EBE test and Schoenfeld-like residual test) had a high level of agreement with the LRT, with the highest level of agreement being observed for the EBE test (99.2% versus 95.4% for Schoenfeld-like residual test). Additionally, the preselection techniques and the LRT had similar power to detect an effect in the 16 scenarios with a "true" covariate effect ($p > 0.05$), as well as the pooled analysis of all 4000 datasets of these 16 scenarios ($p > 0.4$). These findings add to previous research that found comparable power of the EBE test and the LRT to detect covariates in population pharmacokinetic models (5,10). Our findings support the feasibility of preselecting a subset of promising covariates that can be tested with the more computationally intensive LRT. To reduce the risk of failing to preselect a statistical significant covariate, one could use a less stringent significance level during the preselection than that selected for the final selection with the LRT. However, this would also increase the number of preselected false positives, thereby increasing computational time. It is also important to note that the use of preselection techniques does not reduce the importance of considering the scientific plausibility of the tested covariate relationships a priori, as this remains crucial to limit the amount of spurious covariates included (6,17).

In this study, we used a significance level of 5% for all (pre)selection techniques, and therefore expect the observed false positive rate to be around 5%. Both the Schoenfeld-like residual test and the LRT showed a slightly inflated false positive rate (7.2 and 5.7%, respectively) in the pooled analysis of 4000 datasets without a true covariate effect (eight scenarios with 500 repetitions). The false positive rate of the EBE test did not differ significantly from 5% ($p > 0.05$). We did not identify any relationship between the false positive rate and the number of subjects, level of EBE shrinkage or type of covariate. Although EBE shrinkage has been reported to affect the reliability of the EBE test in previous work on population pharmacokinetic models, we did not

Table II. False Positive Rate of (Pre)selection of Binary Covariates [95% CI]

Selection tool	50 subjects		150 subjects	
	HI$_{shrink}$	LO$_{shrink}$	HI$_{shrink}$	LO$_{shrink}$
Schoenfeld	7.0 [5.0–9.7]	7.0 [5.0–9.7]	6.4 [4.5–9.0]	8.4 [6.2–11.3]
EBE test	5.2 [3.5–7.6]	4.8 [3.2–7.2]	5.8 [4.0–8.3]	5.6 [3.8–8.1]
LRT	6.0 [4.2–8.6]	5.2 [3.5–7.6]	5.6 [3.8–8.1]	6.0 [4.2–8.6]

HI$_{shrink}$, high shrinkage scenario ($h_{pop} = 0.005 \, h^{-1}$); LO$_{shrink}$, low shrinkage scenario with ($h_{pop} = 0.05 \, h^{-1}$)

detect any performance issues in the scenarios with high (> 40%) EBE shrinkage in the RTTE models evaluated here (9). These results are similar to what was found by Xu *et al.* for population pharmacokinetic models, and who have even proposed that the EBE test could be used instead of the LRT as a formal covariate selection technique, irrespective of shrinkage, as they provide similar power and the EBE test has improved type I error rate (5).

It is important to recognize that the small inflation of the false positive rate of the Schoenfeld-like residual test will have limited practical impact when it is used as a covariate preselection technique. The inflated false positive rate could lead to a small increase in the number of "false" covariates that are preselected. However, because these covariates will be tested with LRT before final inclusion in the model (Fig. 1), this is therefore unlikely to increase the inclusion of "false" covariates in the final RTTE model. Even with the inflated false positive rate of 7.2%, the Schoenfeld-like residual test filters out the remaining 92.8% of the "truly false" covariates, which can considerably lower the computational time spent performing the LRT in datasets with multiple covariates. As such, the performance of the Schoenfeld-like residual test in covariate preselection can be considered similar to that of the EBE test, although the latter may be more easily applied for time-constant covariates.

One disadvantage of the EBE test is that the evaluation of time-varying covariates can be only conducted via implementing inter-occasion variability. However, the definition of an occasion within a study could be subjective and the implementation requires averaging the covariate values within each occasion. The methodology of the Schoenfeld-like residual test on the other hand is theoretically also suitable for preselection of time-varying covariates: for each event, the expected covariate is re-calculated based on the

covariate values of all uncensored subjects at the time of each event (Eq. 5). Therefore, the calculating the Schoenfeld-like residual does not require covariates to be time-constant. However, it should be noted that we did not include any time-varying covariates in this study, and the performance of the Schoenfeld-like residual test remains to be determined in the future.

Preselection of time-varying covariates would save time, compared to the alternative practice of testing each candidate time-varying covariate using the LRT. The most commonly included time-varying covariate in RTTE models is drug concentration. However, various types of dynamic biomarkers might also be predictive of the between-subject differences in the hazard rate in RTTE models. By including dynamic biomarkers, such as concentration or proteomic or metabolomic biomarker profiles, in RTTE models, we might gain a better understanding of the physiological components that underlie between-subject differences in disease progression and drug effect (18–20). As "omics" techniques typically quantify many compounds in a single platform, the ability to preselect promising covariates in these large datasets becomes more important.

This work expands on existing literature on covariate (pre)selection for pharmacometric RTTE models. Vigan *et al.* have evaluated the Wald test and LRT for binary covariates and found that both techniques resulted in false positive rates close to 5% (2). We found that the LRT has a slightly inflated false positive rate, which only reached significance when all 4000 datasets without a true covariate effect were pooled. An inflated false positive rate of the LRT has also been reported previously in some, but not all, investigations on covariate selection in population pharmacokinetic models (5,10). These different results in different studies might be explained by the fact that the inflation of the false positive rate of the LRT is

Table III. False Positive Rate of (Pre)selection of Continuous Covariates [95% CI]

Selection tool	50 subjects		150 subjects	
	HI$_{shrink}$	LO$_{shrink}$	HI$_{shrink}$	LO$_{shrink}$
Schoenfeld	6.6 [4.7–9.2]	10.2 [7.8–13.3]	5.4 [3.7–7.9]	6.8 [4.8–9.5]
EBE test	3.8 [2.4–6.0]	7.2 [5.2–9.9]	4.8 [3.2–7.2]	4.8 [3.2–7.2]
LRT	5.6 [3.8–8.1]	7.6 [5.5–10.4]	4.6 [3.0–6.9]	5.2 [3.5–7.6]

HI$_{shrink}$, high shrinkage scenario ($h_{pop} = 0.005 \, h^{-1}$); LO$_{shrink}$, low shrinkage scenario with ($h_{pop} = 0.05 \, h^{-1}$)
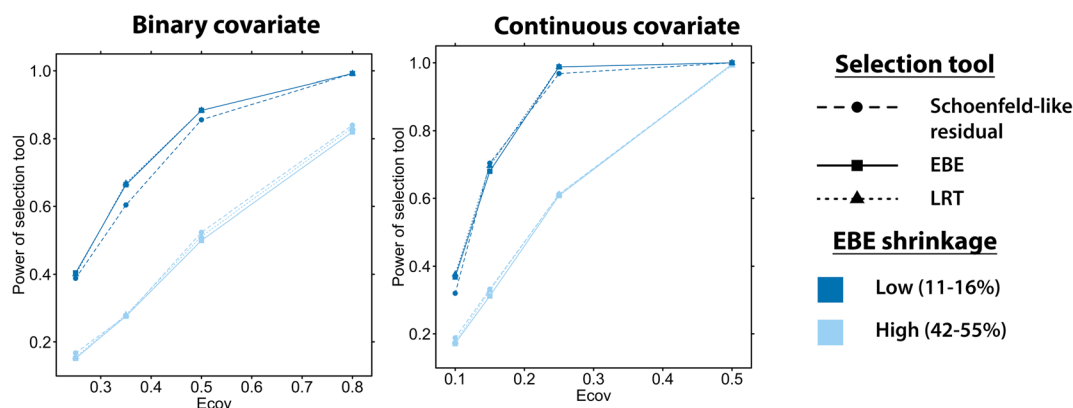
Fig. 2. Power of the different covariate (pre)selection tools to detect a true covariate effect in RTTE datasets. Each scenario was evaluated in 250 simulated datasets of 150 subjects. Low and high shrinkage scenarios were generated by setting the typical hazard to 0.05 or 0.005 h^{-1}, respectively

relatively mild (< 7 with 95% confidence in this study), so that a large number of repetitions is needed to show it to be significantly different from 5%.

In this work, we explored the effect of various factors—shrinkage, number of subjects, type of covariate, covariate effect size—on the performance of covariate (pre)selection techniques. One limitation of this work is that not all potentially relevant scenarios could be explored, such as unbalanced distributions of the binary covariate, time-varying covariates, or datasets with multiple covariates with varying degrees of correlation. The latter scenario could lead to increased false positives when "false" covariates are correlated to "true" covariates (21). Additionally, we did not test highly non-linear covariate relationships.

In this study, we used a fast *R*-script (Supplemental Information) to generate the Schoenfeld-like residual test in situations with time-constant hazard, time-constant covariates, and identical follow-up time for each individual. A more versatile (albeit slower) script can be found in Supplemental Information. This script can be used with time-varying hazard, time-varying covariates, and in the presence of patient dropout. It requires the inclusion of dummy records in the dataset for all non-censored subjects at each time point where one of the subjects experienced an event. These dummy records ensure that for each event, the NONMEM output file includes the population hazard and covariate values during that event for all non-censored subjects (required for Eq. 5 with time-varying population hazard or covariates). The computational time needed for the versatile script depends on the number of subjects and events in the dataset, but was typically below 1 min, and thus considerably faster than the LRT.

CONCLUSION

This study provides the first assessment of the performance of covariate preselection techniques for RTTE models. We found the EBE test to provide a fast and accurate technique for covariate preselection in RTTE models, even in the presence of high EBE shrinkage. The novel Schoenfeld-like residual test proposed in this study has similar performance to the EBE test, and its methodology may be readily applied to time-varying covariates, such as drug concentration and dynamic biomarkers. We also evaluated the false positive

rate of the LRT, the most common method for covariate selection, and found that it has a small but statistically significant inflation of the false positive or type 1 error rate (5.7% instead of 5%) in our RTTE models.

REFERENCES

1. Cox EH, Veyrat-Follet C, Beal SL, Fuseau E, Kenkare S, Sheiner LB. A population pharmacokinetic-pharmacodynamic analysis of repeated measures time-to-event pharmacodynamic responses: the antiemetic effect of ondansetron. J Pharmacokinet Biopharm 1999; 27(6):625–644.
2. 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.
3. 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.
4. Mould DR, Upton RN. Basic concepts in population modeling, simulation, and model-based drug development. CPT Pharmacometrics Syst Pharmacol. 2012;1:e6.
5. Xu XS, Yuan M, Yang H, Feng Y, Xu J, Pinheiro J. Further evaluation of covariate analysis using empirical Bayes estimates

in population pharmacokinetics: the perception of shrinkage and likelihood ratio test. AAPS J. 2017;19(1):264–73.

6. Hutmacher MM, Kowalski KG. Covariate selection in pharmacometric analyses: a review of methods. Br J Clin Pharmacol. 2015;79(1):132–47.

7. Mould DR, Upton RN. Basic concepts in population modeling, simulation, and model-based drug development-part 2: introduction to pharmacokinetic modeling methods. CPT Pharmacometrics Syst Pharmacol. 2013;2:e38.

8. Nguyen TH, Mouksassi MS, Holford N, Al-Huniti N, Freedman I, Hooker AC, et al. Model evaluation of continuous data pharmacometric models: metrics and graphics. CPT Pharmacometrics Syst Pharmacol. 2017;6(2):87–109.

9. Savic RM, Karlsson MO. Importance of shrinkage in empirical Bayes estimates for diagnostics: problems and solutions. AAPS J. 2009;11(3):558–69.

10. Combes FP, Retout S, Frey N, Mentre F. Powers of the likelihood ratio test and the correlation test using empirical Bayes estimates for various shrinkages in population pharmacokinetics. CPT Pharmacometrics Syst Pharmacol. 2014;3:e109.

11. Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika. 1982;69(1):239–41.

12. Collett D. Model checking in the cox regression model. Modelling survival data in medical research. Boca Raton, FL: CRC Press; 2015. p. 131–70.

13. Beal SL, Sheiner LB, Boeckmann AJ, Bauer RJ. NONMEM users guides. 2010.

14. Nyberg J, Karlsson KE, Jönsson S, Simonsson USH, Karlsson MO, Hooker AC. Simulating large time-to-event trials in NONMEM. PAGE. 2014:23.

15. Goulooze SC, Valitalo PA, Knibbe CAJ, Krekels EHJ. Kernel-based visual hazard comparison (kbVHC): a simulation-free diagnostic for parametric repeated-time-to-event models. AAPS J. 2018;20(1):5.

16. Wileyto EP, Li Y, Chen J, Heitjan DF. Assessing the fit of parametric cure models. Biostatistics. 2013;14(2):340–50.

17. Bonate PL. Pharmacokinetic-pharmacodynamic modeling and simulation. 2 ed. Springer, 2011.

18. Kohler I, Hankemeier T, van der Graaf PH, Knibbe CAJ, van Hasselt JGC. Integrating clinical metabolomics-based biomarker discovery and clinical pharmacology to enable precision medicine. Eur J Pharm Sci. 2017;109:S15–21.

19. Kantae V, Krekels EHJ, Esdonk MJV, Lindenburg P, Harms AC, Knibbe CAJ, et al. Integration of pharmacometabolomics with pharmacokinetics and pharmacodynamics: towards personalized drug therapy. Metabolomics. 2017;13(1):9.

20. Goulooze SC, Krekels EHJ, van DM, Tibboel D, van der Graaf PH, Hankemeier T et al. Towards personalized treatment of pain using a quantitative systems pharmacology approach. Eur J Pharm Sci 2017, 109, S32, S38.

21. Therneau TM, Grambsch PM. Functional form. Modeling survival data: extending the cox model. New York: Springer, 2000; p. 87–126.