Application of Resampling Techniques to Estimate Exact Significance Levels for Covariate Selection during Nonlinear Mixed Effects Model Building: Some Inferences

Jogarao V. S. Gobburu[1,3] and John Lawrence[2]

Received July 2, 2001; accepted October 4, 2001

Purpose. One of the main objectives of the nonlinear mixed effects modeling is to provide rational individualized dosing strategies by explaining the interindividual variability using intrinsic and/or extrinsic factors (covariates). The aim of the current study was to evaluate, using computer simulations and real data, methods for estimating the exact significance level for including or excluding a covariate during model building.

Methods. Original data were simulated using a simple one-compartment pharmacokinetic model with (full model) or without (null model) covariates (one or two). The covariate values in the original data were resampled (using either permutations or parametric bootstrap methods) to generate data under the null hypothesis that there is no covariate effect. The original and permuted data were fitted to null and full models, using first-order and first-order condition estimation (with or without interaction) methods in NONMEM, to compare the asymptotic and conditional p-value. Target log-likelihood ratio cutoffs for assessing covariate effects were derived.

Results. The simulations showed that for sparse as well as dense data, the first-order condition estimation methods yielded the best results while the first-order method performs somewhat better for sparse data. Depending on the modeling objective, the appropriate asymptotic p-value can be substituted for the conditional significance level. Target log-likelihood ratio cutoffs should be determined separately for each covariate when exact p-values are important.

Conclusions. Resampling methods can be employed to estimate the exact significance level for including a covariate during nonlinear mixed effects model building. Some reasonable inferences can be drawn for potential application to design future population analyses.

KEY WORDS: mixed effects modeling; covariate selection; hypothesis testing; resampling.

INTRODUCTION

One of the main objectives of drug development is individualization of therapy. Nonlinear mixed effects modeling is widely used to gain insights into pharmacokinetics (PK) and pharmacodynamics (PD) of drugs, using the data typically collected in clinical trials by the regulatory agencies (1) and pharmaceutical companies (2). Towards that end, pharmacometricians attempt to describe the interindividual variability (IIV) of the fundamental PK and PD parameters using meaningful covariates (e.g. relating systemic clearance and body size). Mechanistic reasoning should be the primary criterion for selecting important covariates, and statistical criteria should only be used as supportive evidence. On the contrary, most publications and regulatory submissions report model building based purely on some statistical criteria. Log-likelihood ratios (LLR), determined from the reduced (without covariate) and full models (with covariate), are assumed to follow a chi-square distribution. Thereby, the asymptotic significance level at a specified (not necessarily prospective) Type 1 error probability is calculated.

With the advent of powerful computers, computationally intensive statistical methods are surfacing for use by pharmacometricians. Several resampling methods such as bootstrap permutation, and jackknife techniques are being slowly incorporated into model building activities (3). Each of these methods serves different purpose(s). Bootstrap can be used to determine the confidence regions of the point estimates of model parameters. Permutation technique can be employed to generate the distribution of a given test statistic under null (H0) hypothesis (for example, H0: No covariate effect) (4). Permutation tests are also known as randomization, rerandomization, and exact tests.

The aims of the current project were to (1) investigate the underlying distribution of LLR under null hypothesis and its proximity to the expected chi-square distribution, and (2) demonstrate the application of permutation techniques for estimating exact significance levels for inclusion or exclusion of covariates, for nested models.

METHODS

Simulated Data

Original Data

Data were simulated using a simple one-compartment model with a typical systemic clearance (CL_{typ}) of 5 L/h and a typical volume of distribution (V_{typ}) of 50 L. The IIV of all parameters were assumed to follow a log-normal distribution with a 30% coefficient of variation (CV). The residual variability was assumed to follow a combined proportional (10% CV) and additive error model (standard deviation of 0.1 mg/L). Data were simulated without (NULLMODEL) covariate effects.

Whereas mechanistic modeling is the most rationale method of constructing models, for the sake of investigating the statistical properties, we intentionally chose to make certain empirical simplifications. In the experiments that included effects of body weight (on CL) and age (on V), the following models were used:

$$CL = CL_{typ} + \text{wtcl} \cdot (\text{WT-70}) \quad (1)$$

$$V = V_{typ} + \text{agev} \cdot (\text{AGE-50}) \quad (2)$$

Where wtcl (0.1 L/h/kg) and agev (0.1 L/y) are the slopes of the linear relationships between the PK parameter and the

[1] Pharmacometrics, Division of Pharmaceutical Evaluation—1, Office of Clinical Pharmacology and Biopharmaceutics, Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland.

[2] Biometrics, Office of Blostatistics, Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland.

[3] To whom correspondence should be addressed 1451 Rockville Pike, Room 5088, HFD-860, Rockville, Maryland 20852. (e-mail: gobburuj@cder.fda.gov)

[4] The views expressed in this article are those of the authors and do not reflect the official policy of the FDA. No official endorsement by the FDA is intended or should be inferred.

0724-8741/02/0100-0092/0 © 2002 Plenum Publishing Corporation

covariate. The data were centered at a typical WT of 70 kg and an AGE of 50 years. The WT and AGE were simulated using a log-normal distribution with a mean of 70 kg and 50 years and IIV with a CV of 30%, respectively. Further, we assumed a covariance between AGE and WT with a correlation coefficient of 60%. The model that assumed only WT to be an influential covariate will be referred to as WTMODEL (Eq. [1]), the model that assumed only AGE to be influential will be referred to as AGEMODEL (Eq. [2]) and the model that assumed both WT and AGE to be influential will be referred to as WTAGEMODEL (Eqs. [1 and 2]).

Data under dense (0.25, 2, 6, 12, 24 h) and sparse (0.25, 12 h) sampling designs were simulated with 30 subjects. Mixed sampling design included 50% dense and 50% sparse sampling. To explore the underlying LLR distribution, 1000 replications of original data (OD) were simulated using the NULLMODEL.

Permuted Data

The LLR ratio is intended to measure how much better the full model describes the data compared to the null model. Large values of the LLR tell us that the full model describes the data better. LLRs obtained using NONMEM are assumed to follow a chi-square distribution (5). To determine how large the observed value must be to convince us to use the full model, we need to find out how the LLR is distributed when the reduced model is, in fact, true. By permuting the covariates, we can simulate a data set where the covariates are guaranteed to be unrelated to the model parameters. By repeatedly doing this, we can estimate the distribution of the LLR under the null model.

Rerandomized covariates can be obtained in of the following three ways: (1) by permuting the raw covariates (6), (2) by permuting the residuals under the full model (7), and (3) permuting the residuals under the reduced model obtained by first correcting for the covariate(s) (8). We employed the first of these methods in our analyses because of its ease of implementation whose details are provided below.

The covariates in OD were permuted (1000 times) to simulate (PermData) data under the null hypothesis that wtcl = 0. For instance, when the true model was WTMODEL, under null hypothesis the WTs do not contribute significantly to decreasing the IIV, and it does not matter whose WT is what. Such Monte–Carlo simulations can be performed in a parametric (bootstrap) or a nonparametric (permutations or bootstrap) manner. Nonparametric permutations, which is resampling without replacement, were achieved by randomly swapping the covariate values among the subjects. Bootstrap, which is resampling with replacement, could also be employed to simulate under null hypothesis. Parametric permutations (better known as ‘parametric bootstrap’) were achieved by randomly simulating the covariate using a parametric distribution. The number of permutation replications was determined using the binomial theorem:

$$\mathrm{Nrep} = \frac{\mathrm{p} \cdot (1 - \mathrm{p})}{\mathrm{SE}^2} \quad (3)$$

Where p is the probability of an event of interest, SE is the target standard error, and Nrep is the number of replications. For example, if the alpha level of interest is 5% (p = 0.05) and target SE is 0.007 (14% of p) then Nrep would be 1000.

Log-Likelihood Ratio (LLR) Distribution

The aim of these experiments was to explore the LLR distribution obtained for each of the estimation methods. The OD and PermData (Nrep = 1000) were fitted to NULLMODEL and WTMODEL. The difference in the objective function value between the two models, i.e. the LLR, was calculated. Kolmogorov-Smirnov (KS) nonparametric statistical testing was used to either accept or reject the hypothesis that the LLR followed a chi-square distribution. The probabilities of the LLRs being within 3.84 (α 5%), 6.63 (α 1%), and 10.83 (α 0.1%) were also determined.

Estimation of Significance Level

The aim of these experiments was to conduct a thorough comparison of the asymptotic and conditional (via permutations) p-values. To that end, 100 ODs were randomly simulated using the NULLMODEL. True (NULLMODEL) and alternate (WTMODEL) models were fitted to each of the OD sets to determine the asymptotic p-value. It is prohibitively cumbersome to permute 100 ODs for 1000 times, thus we permuted 100 times instead. The asymptotic p-value was determined assuming that the LLR followed a chi-square distribution with one degree of freedom. Further, parametric permutations (Nrep = 100) of the WTs were performed and the alternate model was fitted to the permuted data to determine the conditional p-value. The conditional p-value was generated by counting the number of times the LLR of the permuted data was greater than 3.84.

In a different set of experiments, we evaluated the question of whether a particular target LLR cutoff can be universally used for all covariates or not, given a data set. Dense data were simulated using WTAGEMODEL and fitted to NULLMODEL, WTMODEL, AGEMODEL, and WTAGEMODEL. Permuting AGE among the patients without considering WT will produce considerable bias into the estimation under the null hypothesis. For example, it might be unreasonable to allow a 100 kg patient to be one year old. The functional relationship between WT and AGE was described using a linear model and the residual error was permuted among the subjects. The permuted residual error was added to the AGE predicted by the so-developed linear model to randomly generate AGE (PermData), given the (true) WT of the subject. Target LLR cutoffs, at an alpha level of 5%, were derived for WT (WTMODEL) and AGE (AGEMODEL) separately as well as for AGE given WT (WTAGEMODEL) already in the model, using nonparametric permutations. The LLR from the 1000 permuations were ordered, and the value at the 95th percentile is identified as the target LLR cutoff for an alpha of 5%.

Real Data

The purpose of using real data was to demonstrate the application of resampling techniques in selecting covariates. Hence the identity of the drug *per se* is not relevant to achieve this objective. The data are from a clinical trial in which an oral antiepileptic drug was administered b.i.d., to about 100 patients with ages between 3 and 17 years and body weights between 16 and 100 kg. About 50% of the patients received another drug-drug interaction (DDI) concomitantly, which potentially could interact with the pharmacokinetics of the drug of interest. The presence or absence of the concomitant medication was indicated by 1 and 0 (zero) in the data set. Plasma concentrations (about 3.5 samples/patient) were obtained at steady state. The PK were described using a simple one-compartment model with a first-order absorption. From decades of experience, body size is known to be an important predictor of CL and V, according to the allometric model:

$$CL = CL_{typ} \cdot (WT/70)^{beta} \quad (4)$$

$$V = V_{typ} \cdot (WT/70) \quad (5)$$

Where, CL_{typ} is the typical clearance in a 70 kg person, beta is the exponent, and V_{typ} is the typical volume of distribution in a 70 kg person. The PK model with body weight as a covariate was employed as the base model. The residual error was described using a combined proportional and additive error model. Subsequent analysis aimed at answering the questions: (1) Is age an important predictor of clearance?, and (2) Does the concomitant administration of DDI alter the clearance of the drug of interest? The influence of age on CL was tested using

$$CL = CL_{typ} \cdot (WT/70)^{beta} \cdot [1 + agecl \cdot (AGE\text{-}10)] \quad (6)$$

The influence of DDI on CL was tested using

$$CL = CL_{typ} \cdot (WT/70)^{beta} \cdot (1 + ddicl \cdot DDI) \quad (7)$$

In Eqs. (6) and (7), agecl is the slope of the age and CL relationship and ddicl is the fraction by which CL changes in the presence of DDI. To test if ddicl is significantly different from zero, the ones and zeroes of DDI were randomly swapped among the patients to simulate PermData. This ensures that the CL estimate is independent of the fact that the patient received DDI or not. Taking the range of ages and body weights into consideration, *a priori* we can expect that WT and AGE are not mutually independent. Permuting AGE among the patients without considering WT will produce considerable bias into the estimation under the null hypothesis. The functional relationship between WT and AGE was described using a linear model and the residual error was permuted among the subjects. The permuted residual error was added to the AGE predicted by the so-developed linear model to randomly generate AGE (PermData), given the (true) WT of the subject. The original (real) and PermData (Nrep = 1000) were fitted to base model with or without the covariate (equation 6 or 7). The difference in the objective function value (OBJ) between the two models, i.e. the LLR, was calculated. Quantile–Quantile (QQ) plots were used to assess whether the data have a chi-square distribution with one degree of freedom. If the distributions are the same, then the plot will be approximately a straight line. Kolmogorov-Smirnov (KS) nonparametric statistical testing was used to either accept or reject the hypothesis that the LLR followed a chi-square distribution. The 95th percentile of the observed LLR distribution (using PermData) was determined.

All data were simulated and modeled using NONMEM (ver. 5.0, level 1.1) with a Compaq Digital Fortran compiler (ver. 6.01). The model parameters were estimated using first-order (FO) and first-order conditional without (FOCE), or with interaction between inter-individual and residual errors (FOCE-INTER) estimation methods provided in NONMEM. Permutations and other data manipulations were performed using SAS (Cary, North Carolina) (ver. 6.12) and S-plus 2000.

RESULTS AND DISCUSSION

Simulated Data: LLR Distribution

Figure 1 shows the distribution of theoretical and observed LLRs when the data are dense. By visual inspection, the LLR distributions estimated using the FOCE and FOCE-INTER methods are in good agreement with the theoretical asymptotic distribution and with most values less than 8. The LLR distribution estimated using the FO method seems to be more "spread" out. For example, only 60% of the LLRs are less than or equal to 2 for the FO method, as opposed to about 80% for the FOCE methods. Fitting dense data using the FO method, then, would lead to overestimation of the significance level (or more false positives). Table I shows the probabilities of observed LLRs for the 3 estimation methods. Whereas the expected probability of LLRs to be greater than or equal to 3.84 is 0.05 at an alpha of 5%, the observed probability was 0.17. The corresponding probabilities for the FOCE and FOCE-INTER are 0.078 and 0.065, which are in better agreement with 0.05. The trend is similar at all other alpha levels tested. The KS test showed that the distribution of LLRs derived using the FO method are significantly different (p-value ~0) from the theoretical chi-square distribution with one degree of freedom. FOCE-INTER method yielded LLRs not significantly different from an expected chi-square distribution, based on the KS test.

When the data are sparse, the FO method results in probabilities relatively closer to the expected values, although not better than FOCE or FOCE-INTER. Although the LLR distribution was found to be very significantly different from the chi-square distribution with one degree of freedom, from a practical consideration, one can employ the FO method to derive reasonable asymptotic p-values. The contribution of the earlier part of the LLR probability distribution curve toward the cumulative probability is much larger than that of the later part. Hence, it is possible to find significant deviations from the chi-square distribution, yet, have p-values relatively close to the alpha level. The p-values for rejecting the null hypothesis that the observed LLR distribution follow

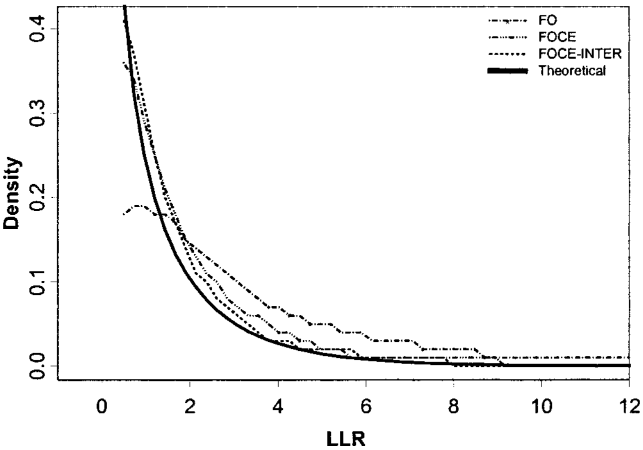

Fig. 1. Theoretical and observed distributions of the log = likelihood ratios. Simulations included 30 subjects and a dense sampling schedule.

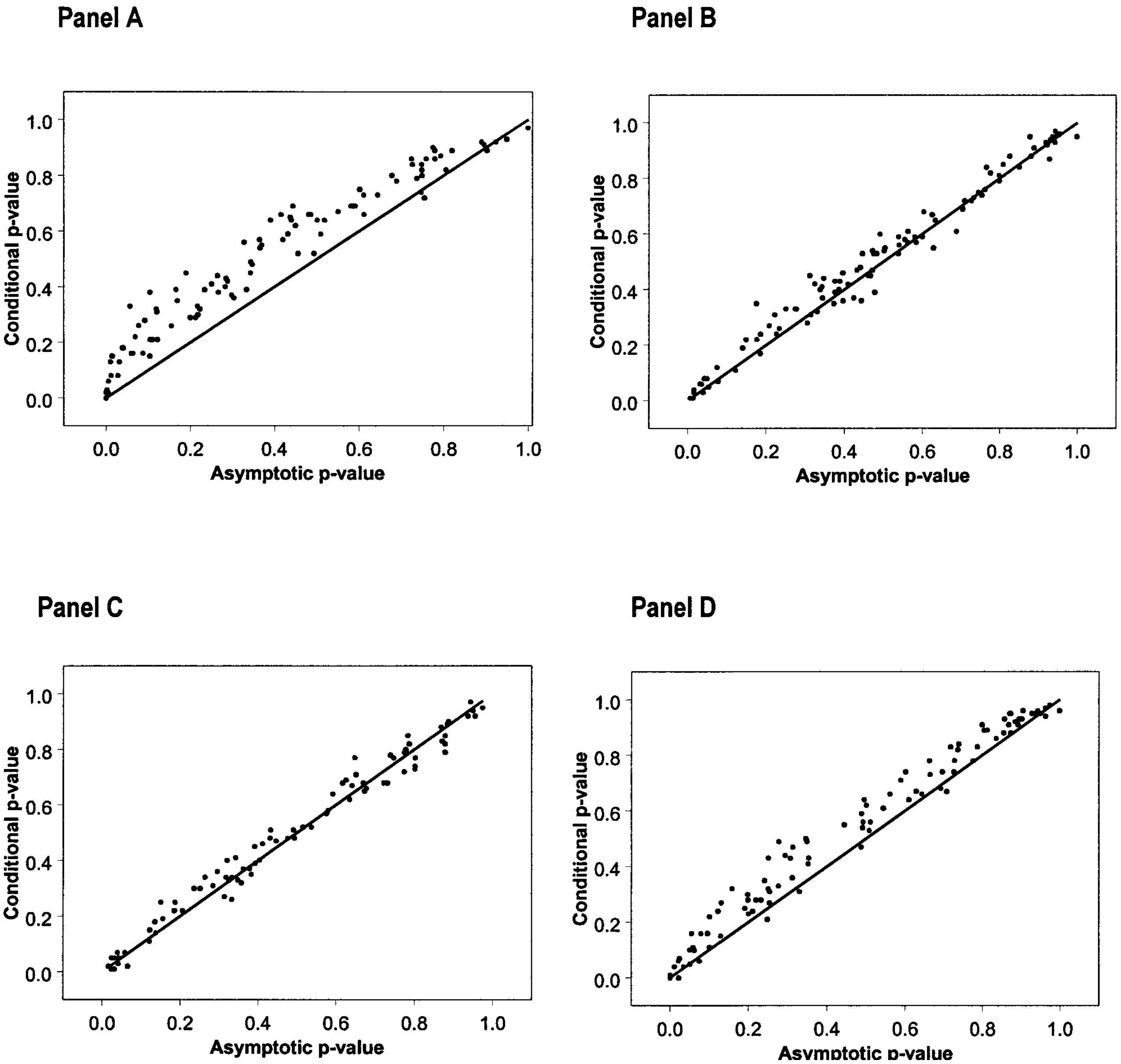

Fig. 2. The asymptotic and conditional p-values derived by fitting the data simulated using the NULLMODEL to NULLMODEL and WTMODEL. Panel A, B, and C show the results from using FO, FOCE, and FOCE-INTER estimation methods, respectively, when the sampling was dense. Panel D shows the results when the FO method was used with sparse data. The solid line represents the line of identity.

chi-square distribution are large for FOCE and FOCE-INTER estimation methods. Inferences for the case when the data are partly dense and partly sparse are similar to those for the dense data case. Hence, when conducting a meta-analysis of data from many clinical trials (mixture of sparse and dense data) the appropriate FOCE (depending on the residual error model) is the preferred method.

Simulated Data: Estimation of Significance Level

Figure 2 shows the asymptotic and conditional p-values derived using the FO, FOCE and FOCE-INTER methods for dense data. The FO method offers poor asymptotic "estimates" of the significance levels that deviate systematically from the line of identity (Panel A). The asymptotic p-values are under-predicted (i.e. more significant than they really are) as evident from the convex shape of the scatter plot. For example, an asymptotic p-value of 0.05 estimated using the FO method corresponds truly to a conditional p-value of 0.18. Based on the statistical significance, one could falsely conclude that the covariate WT significantly contributes to describing the interindividual error. Even when that data are sparse, the FO method under-predicted the significance level, but not as poorly as for the dense data. This is not the case when the FOCE estimation methods are used for both dense and sparse data (not shown). The significance levels distribute evenly around the line of identity. Although FOCE with or without interaction yielded very similar results, this should not be interpreted as a general case. The FOCE-INTER method was built to allow interaction between the two sources of random variability (inter-individual and residual errors).

Table I. Probabilities of Observed Log-Likelihood Ratios at 5%, 1%, and 0.1% Alpha Levels

Probability	FO	FOCE	FOCE-INTER
Dense data			
α5%	0.170	0.078	0.065
α1%	0.077	0.019	0.014
α0.1%	0.020	0.004	0.002
p-value	0	0.0017	0.241
Sparse data			
α5%	0.079	0.054	0.057
α1%	0.025	0.014	0.014
α0.1%	0.004	0.002	0.002
p-value	0	0.305	0.122
Mixed data			
α5%	0.132	0.073	0.064
α1%	0.045	0.02	0.013
α0.1%	0.012	0.004	0.002
p-value	0	0.204	0.155

Note: The probability of the LLR to be less than 3.84 (based on a chi-square distribution with 1 degree of freedom at a 5% alpha level) is about 0.17, when the expected expected value is 0.05. The p-value from the KS test for rejecting the null hypothesis that the observed distribution is not significantly different from a chi-square distribution, with one degree of freedom, is also provided).

Further experiments to investigate the applicability of a universal target LLR cutoff (for an α = 5%) showed that target LLR cutoffs need to be derived separately for every covariate. The target LLR cutoffs are shown in Table II. Clearly, the target LLR cutoffs are different for different covariates and are also sensitive to the covariates already present in the model. The target LLR for considering WT in the model is 15.37 whereas that for the AGE is 9.72. Further, the target LLR for considering AGE when WT is already present in the model is 6.46. The target LLRs get closer for the FOCE methods, especially for FOCE-INTER, but not identical. This is expected from the fact that the closer the LLR distribution resembles the chi-square distribution, the closer the target LLR cutoff would be to the theoretically expected value. It should be noted, however, that one would naturally use the FO method due to practical constraints on the execution time, if deriving exact p-value were critical.

Table II. Simulation Data: The Asymptotic and Conditional *p*-Values Estimated Using FO, FOCE, and FOCE-INTER Methods

Covariate	LLR cutoff	Asymptotic p-value	Conditional p-value
FO			
WT	15.37	0.000	0.000
AGE	9.72	0.964	0.986
AGE|WT	6.46	0.020	0.069
FOCE			
WT	5.59	0.000	0.000
AGE	5.28	0.000	0.001
AGE/WT	5.67	0.199	0.284
FOCE-INTER			
WT	3.94	0.000	0.000
AGE	3.85	0.343	0.360
AGE|WT	4.29	0.316	0.331

Note: The log-likelihood ratio cut-off values derived based on the 95th percentile are also shown for an α = 5%, for WTMODEL (WT), AGEMODEL (AGE), and WTAGEMODEL (AGE|WT). The exepected cutoff value according to the chi-square distribution is 3.84 for an α = 5% and for one degree of freedom.

Real Data: Estimation of Significance Level

The base model for analyzing the real data included body weight as a covariate. Age was added to the base model to describe clearance using Eq. (6). Permuting the ages conditional upon the body weight of a given patient was performed to derive the exact p-value for including age as a covariate. For instance, it may be meaningless to randomly assign a WT of 70 kg to a neonate. The relationship between body weight and age of the patients was described using a linear function to perform meaningful simulations. Figure 3 shows the QQ plots when age was added as a covariate to the base model for both FO and FOCE-INTER estimation methods. For the FO estimation method, the QQ plot suggests that the LLR distribution deviates from the expected chi square distribution. The QQ plot for the FOCE-INTER estimation method demonstrates that the LLR distribution is reasonably similar to the expected chi-square distribution. However, the KS test indicated that the observed LLR distribution is significantly different from a chi-square distribution with one degree of freedom for both of the estimation methods (α = 5%). Table III shows the asymptotic and conditional p-values for both estimation methods. The LLR cutoff value for the FO method was found to be 6.99 as against the expected 3.84 (α = 5%). The LLR cutoff value for the FOCE-INTER method (4.67) was more reasonable than that of the FO method. The FO method, although much greater than 0.05, overestimated the significance level of age (0.2490 vs. 0.3980). These results suggest that one could employ the FO method to determine the significance of including age by first estimating the target LLR cutoff via permutations. However, one can circumvent the grueling task of permuting and fitting several hundreds of replications by using the FOCE-INTER method that offers more reliable asymptotic p-values (0.4511 vs. 0.5180). Results from the simulated data, as shown in Fig. 2 (Panel A), also

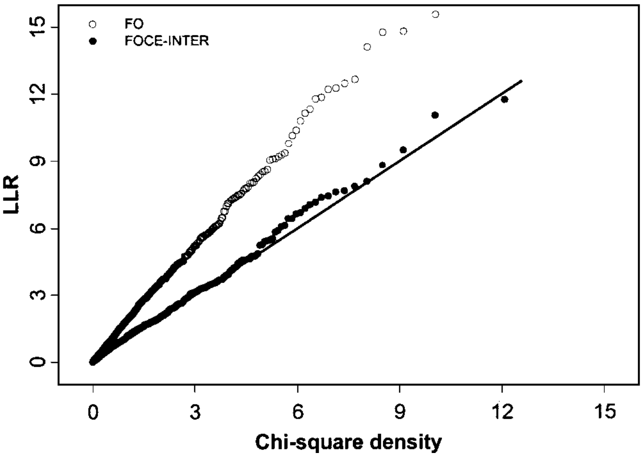

Fig. 3. QQ plots to assess the influence of FO (hollow circles) and FOCE-INTER (filled circles) estimation method deriving asymptotic and conditional p-values when AGE was included in the model. The solid line represents the line of identity.

demonstrate that the asymptotic p-value determined using the FO method is "anticonservative." Nevertheless, both estimation methods suggested that age is not an important determinant of clearance.

Permuting the DDIs was performed to derive the exact p-value for including DDI as a covariate. QQ plots when DDI was added as a covariate to the base model, for both FO and FOCE-INTER estimation methods, were similar to those shown in Fig. 3. For the FO estimation method, the QQ plots suggested that the LLR distribution deviates from the expected chi-square distribution. The QQ plot for the FOCE-INTER estimation method demonstrated that the LLR distribution is reasonably similar to the expected chi-square distribution. However, the KS test indicated that the observed LLR distribution is significantly different from a chi-square distribution with one degree of freedom for the both of the estimation methods (alpha = 5%). Table III shows the asymptotic and conditional p-values for both estimation methods. The LLR cutoff value for the FO method was found to be 6.48 as against the expected 3.84 (α = 5%). The LLR cutoff value for the FOCE-INTER method (3.70) was more reasonable than that of the FO method. As mentioned earlier, one could utilize permutations to establish the target LLR cutoff using the FO method. But a more pragmatic solution would be the use of FOCE-INTER to estimate the asymptotic p-value that is reasonably close to the conditional p-value. Nevertheless, both methods suggested that DDI is an important determinant of clearance. We employed nonparametric permutations to determine the exact p-values. It is also possible to employ parametric bootstrap techniques as an alternative. No assumption about the functional form of the distribution is required when using nonparametric permutations. Further, if nonparametric permutations are performed, it is relatively easier to deal with time-varying covariates.

Recently, Wählby *et al.* reported the use of permutations to screen covariates using a wide variety of models (9). The results presented in the current article, by and large, are in agreement with Wählby *et al.* findings. In their work, the authors simulated a random binary variable (dummy covariate) to derive the target LLR cutoff value for a given dataset that was suggested to be valid for all covadates. Our results and theoretical expectations seem to suggest differently. Evidently, the cutoff values for age and DDI are different, particularly for the FO method. The simulated data illustrate this point even more clearly. Even Wählby *et al.* (9) obtained different actual significance levels corresponding to an alpha of 5% and 1% for different covariate characteristics. These authors concluded that one target LLR cutoff fits all covariates. For example, Wählby *et al.* reported that the actual significance level at an alpha of 1% was 7.89 and 10.66, for the cases where a binary covariate (0 or 1) was simulated with a ratio of 10:90% and 2:98%. These conditional cutoff values are clearly different. Banken (10) reported a similar trend when the binary covariate ratio was varied from 5% to 50%. The underlying distributions of these covariates are different and hence one should expect different cutoff values. Further, when handling covariates like serum creatinine, body weight, age, and gender, we strongly feel their correlation needs to be taken into consideration to derive reliable significance levels.

Table III. Real Data: The Asymptotic and Conditional p-Values Estimated Using FO and FOCE-INTER Methods

Covariate	Estimation method	AsymP[a]	CondP[b]	Cut-off
Age	FO	0.249	0.398	6.99
Age	FOCE-INTER	0.451	0.518	4.67
DIDI	FO	0.000	0.005	6.48
DDI	FOCE-INTER	0.007	0.010	3.70

[a] The AsymP (asymptotic p-value) was determined assuming that the LLR followed a chi-square distribution with one degree of freedom.

[b] The CondP (conditional p-value) was generated by counting the number of times the LLR of the permuted data was greater than 3.84.

Note: The log-likelihood ratio cut-off values derived based on the 95th percentile are also shown for an α = 5%. The expected cut-off value according to the chi-square distribution is 3.84 for an α = 5% and for one degree of freedom.

Monte-Carlo simulations, such as permutations, offer a good means of estimating the exact p-value during nonlinear mixed effects model building. However, the disadvantage is that of the run time. It is well appreciated that model building is highly time demanding even for moderately big datasets (which are most common). Under conditions utilized to generate the results in Table II, the FO method required 7 h and the FOCE methods required about 35 h on an 800 MHz Pentium computer with 128 MB RAM (Compaq). It is important to note that the dataset and the model are fairly simple in our case. If the FO method requires 1 h for each run and there are 10 models to be evaluated with a p-value precision of about 15% (Nrep = 1000), then the total modeling time would be 10,000 h.

CONCLUSIONS

Monte–Carlo simulations, such as permutations, offer a good means of estimating the exact p-value during nonlinear mixed effects model building. However, the disadvantage is that of an increase in time, which is not trivial. It is impractical to estimate model parameters of simulated data over several hundreds of replications using the FO or FOCE method on a day-to-day basis. Identification of covariates needs some level of mechanistic understanding and *a priori* expectation. When such understanding is available deriving the exact p-value may not be required and hence tedious simulations via permutations may not be necessary. However, an appropriate estimation method has to be selected for a given data set (see below). Developers should incorporate methods, such as the one described in the present report, for qualifying a model in the software packages for wider and ready availability. When the exact p-value is required, for example to establish the effectiveness of a new molecular entity, permutations may be necessary.

Our results suggest the following:

1. When at least some subjects have dense sampling, the FO method offers a poorer estimate of the significance level. The FOCE-INTER method should be used. Again, if exact p-value is required then permutations and estimation using the FO method is recommended.

2. Even when the data are sparse the FOCE methods perform the best. However, the asymptotic p-values derived using the FO method are more reasonable for the sparse data than those for the dense data. The LLR distribution in both

cases is significantly different from the chi-square distribution. If an exact p-value is required, then permutations and estimation using the FO method is recommended.

3. The LLR cutoff value should be determined separately for each covariate. One cutoff value does not apply for all covariates, especially with the FO method. The FOCE methods provide reasonably close target cutoff values for all covariates. The correlation between the covariates already in the model and the covariates under testing should be taken into account.

4. In general, good mechanistic reasoning would efficiently avoid the practical limitations of the extremely time demanding task of simulation and estimation to determine exact p-values.

Modeling is more of an art rather than a fully developed science. Whereas certain components of the model building process cannot be formalized, some stages could be. One of the most important components would be that of model qualification (11). The analysts and others in the drug development team (including regulators) need to ascertain themselves that the proposed model(s) are qualified to achieve the purpose for which they are built. False positive significance levels may unnecessarily increase the cost of drug development and complicate the drug labeling. At the same time, not being able to recognize important covariates might lead to suboptimal use of a drug. The current and previous reports offer useful guidelines in using permutation tests to select covariates during model development. We would like to encourage researchers to devote efforts to construct good modeling practices, which we believe would enhance the credibility of the applications of modeling and simulation in contemporary drug development. Future research should focus on obtaining exact p-values for non-nested models, evaluating various resampling techniques to handle time-varying covariates, and how to deal with run failures estimation of permuted samples.

REFERENCES

1. Guidance for Industry: Population Pharmacokinetics. Center for Drug Evaluation and Research, United States Food and Drug Administration, http://www.fda.gov/cder/guidance/index.htm (1999).
2. E. Samara and R. Granneman. Role of population pharmacokinetics in drug development. A pharmaceutical industry perspective. *Clin Pharmacokinet.* **32**:294–312 (1997).
3. B. Efron and Tibshirani R. J. An introduction to the bootstrap. *Chapman & Hill* New York (1993).
4. Good P. Permutation tests: A practical guide to resampling methods for testing hypotheses. *Springer,* 2nd ed., 2000.
5. S. L. Beal and L. B. Sheiner (eds). NONMEM user's guides. San Francisco, CA: NONMEM Project Group, University of California (1992).
6. M. J. Anderson and P. Legendre. An empirical comparison of permuatation methods for tests of partial regression coefficients in a linear model. *J. Statist. Comp. Simul.* **62**:271–303 (1999).
7. C. J. F. ter Braak. Permuation versus bootstrap significance tests in multiple regression and ANOVA. In: K. H. Jockel, G. Rothe, and W. Sendler (eds), *Bootstrapping and related techniques.* Springer-Verlag, Berlin, pp. 79–82 (1992).
8. D. Freedman and D. Lane. Nonstochastic interpretation of reported significance levels. *J. Bus. Econ. Stat.* 1:292–298 (1983).
9. U. Wählby, E. N. Jonsson, and M. O. Karlsson. Assessment of actual significance levels for covariate effects in NONMEM. *J. Pharmacokinet. Pharmacodyn.* **28**:231–252 (2000).
10. L. Banken. Evaluating the properties of statistical tests through simulations. *Presented at the Basel Biometric Society (BBS) Meeting on Resampling and Simulation in Medical Research,* June 6, 2001.
11. J. V. S. Gobburu and P. J. Marroum. Utilization of PK/PD modeling and simulation in regulatory considerations. *Clin. Pharmacokinet.* 40:883–892 (2001).