Published in final edited form as:
J Pharmacokinet Pharmacodyn. 2013 April ; 40(2): 243–252. doi:10.1007/s10928-013-9313-5.

Cross-Validation for Nonlinear Mixed Effects Models

Emily Colby and
Dept. of Biostatistics, Univ. of North Carolina-Chapel Hill, Chapel Hill, NC, ecanimation@hotmail.com

Eric Bair
Depts. of Endodontics and Biostatistics, Univ. of North Carolina-Chapel Hill, Chapel Hill, NC, Tel.: 919-537-3276, Fax: 919-966-5339, ebair@email.unc.edu

Abstract

Cross-validation is frequently used for model selection in a variety of applications. However, it is difficult to apply cross-validation to mixed effects models (including nonlinear mixed effects models or NLME models) due to the fact that cross-validation requires "out-of-sample" predictions of the outcome variable, which cannot be easily calculated when random effects are present. We describe two novel variants of cross-validation that can be applied to nonlinear mixed effects models. One variant, where out-of-sample predictions are based on post hoc estimates of the random effects, can be used to select the overall structural model. Another variant, where cross-validation seeks to minimize the estimated random effects rather than the estimated residuals, can be used to select covariates to include in the model. We show that these methods produce accurate results in a variety of simulated data sets and apply them to two publicly available population pharmacokinetic data sets.

Keywords

cross-validation; model selection; nonlinear mixed effects; population pharmacokinetic modeling

1 Introduction

1.1 Overview of Population Pharmacokinetic and Pharmacodynamic Modeling

Population pharmacokinetic and pharmacodynamic (PK/PD) modeling is the characterization of the distribution of probable PK/PD outcomes (parameters, concentrations, responses, etc) in a population of interest. These models consist of fixed and random effects. The fixed effects describe the relationship between explanatory variables (such as age, body weight, or gender) and pharmacokinetic outcomes (such as the concentration of a drug). The random effects quantify variation in PK/PD outcomes from individual to individual.

Population PK/PD models are hierarchical. There is a model for the individual, a model for the population, and a model for the residual error. The individual PK model typically consists of a compartmental model of the curve of drug concentrations over time. The pharmacokinetic compartmental model is similar to a black box engineering model. Each of the compartments is like a black box, where a system of differential equations is derived based on the law of conservation of mass (Sandler, 2006). The number of such compartments to include in the model must be determined based on the data.

The equations for the PK/PD parameters represent the model for the population in the hierarchy of models. The PK/PD parameters are modeled with regression equations containing fixed effects, covariates, and random effects (denoted by η's). The random effects account for the variability across subjects in the parameters and for anything left out of the parameter equations (such as a covariate not included). The vector of random effects (η) is assumed to follow a multivariate normal distribution with mean 0 and variance-covariance matrix Ω. The matrix Ω may be diagonal, full block, or block diagonal. The model for the residual error (ε) accounts for any deviation from the model in the data not absorbed by the other random effects. The residual error model may be specified such that measurements with higher values are given less importance compared with measurements with smaller values, often referred to as "weighting".

Hence, population PK/PD models are non-linear mixed effects (NLME) models. They are represented by differential equations that may or may not have closed-form solutions, and are solved either analytically or numerically. The parameters are estimated using one of the various algorithms available such as first order conditional estimation with interaction (FOCEI). See Wang (2007) for a mathematical description of these algorithms.

Once model parameters are estimated using an algorithm such as FOCEI, one may fix the values of the model estimates and perform a post hoc calculation to obtain random effect values (η's) for each subject. Thus, one may fit a model to a subset of the data and obtain random effect values for the full data set. See Wang (2007) for a discussion of how these posterior Bayes (post hoc) estimates of the η's are calculated.

1.2 Cross-Validation and Nonlinear Mixed Effects Modeling

In general, cross-validation is not frequently used for evaluating nonlinear mixed effects (NLME) models (Brendel et al, 2007). When cross-validation is applied to NLME models, it is generally used to evaluate the predictive performance of a model that was selected using other methods. For example, in Bailey et al (1996), data were pooled across subjects to fit a model as though the data were obtained from a single subject. Subjects were removed one at a time, and the accuracy of the predicted observations with subsets of the data was assessed. Another approach (Hooker et al, 2008) removed subjects one at a time to estimate model parameters and predicted PK parameters using the covariate values for the subject removed. The parameters were compared with the PK parameters obtained using the full data set in order to evaluate the final model and identify influential individuals. See Mulla et al (2003), Kerbusch et al (2001), and Rajagopalan and Gastonguay (2003) for additional examples where cross-validation was used to validate NLME models.

Less frequently cross-validation is used for model selection in NLME modeling. For example, one may wish to compare a model with a covariate to another model without the covariate. In Ralph et al (2006), the prediction error of the posterior PK parameter for each subject was calculated, and a paired t-test was performed to compare the prediction error between a base and full model to assess whether differences between the models were significant. The full model was only found to be correct when the effect of the covariate was large.

In several published studies, cross-validation failed to identify covariate effects that were identified using other methods. As noted earlier, Ralph et al (2006) found that cross-validation only identified covariate effects when the effect was large. Similarly, Zomorodi et al (1998) found that cross-validation tended to favor a base model (without a covariate) despite the fact that the covariate was found to be significant using alternative approaches. Fiset et al (1995) also found that models with and without covariates tended to produce comparable error rates despite the fact that likelihood-based approaches favored models that included covariates. Indeed, Wahlby et al (2001) used a special form of cross-validation where one concentration data point was chosen for each parameter, which was the point at which the parameter was most sensitive based on partial derivatives. Once again, little difference was observed between models that included covariates and corresponding models without covariates. Thus, cross-validation can fail to detect covariate effects even when attention is restricted to a subset of the data that should be most sensitive to model misspecification.

Despite the fact that cross-validation may fail to detect covariate effects, it has been successfully used to compare models with structural differences, such as a parallel Michaelis-Menten and first-order elimination (MM+FO) model and a Michaelis-Menten (MM) model (Valodia et al, 2000). This indicates that cross-validation can be used for model selection in NLME modeling under certain circumstances. Moreover, the fact that cross-validation often fails to detect covariate effects is not surprising. When covariate effects are present in an underlying NLME model, a misspecified model that fails to include a covariate may not significantly decrease the predictive accuracy of the model. This can occur when random effects in the pharmacokinetic parameters can compensate for the missing covariate. Thus, if cross-validation chooses the model with the lowest out-of-sample prediction error, it may not be able to determine whether a covariate should be included in the model.

Other methods have been proposed for using cross-validation for model selection in NLME modeling (Ribbing and Jonsson, 2001; Katsube et al, 2011). However, these methods rely on estimation of the likelihood function, which is unusual for cross-validation, and they have not been studied extensively.

Thus, we propose an alternative form of cross-validation for covariate model selection in NLME modeling. Rather than choosing a model which minimizes the out-of-sample prediction error, we choose a model which minimizes the post hoc estimates of the random effects (η's). The motivation is that if the η's are large, this suggests that there is a large amount of unexplained variation from individual to individual, which indicates that a covariate may be missing from the model. However, traditional cross-validation (which minimizes the out-of-sample prediction error) is still useful for comparing structural models, as we will discuss below.

2 Methods

2.1 Cross-Validation

Cross-validation is a method for evaluating the expected accuracy of a predictive model. Suppose we have a response variable Y and a predictor variable X and we seek to estimate Y based on X. Using the observed X's and Y's we may estimate a function $\widehat{f}$ such that our estimated value of Y (which we call $\widehat{Y}$) is equal to $\widehat{f}(X)$. Cross-validation is an estimate of the expected loss function for estimating Y based on $\widehat{f}(X)$. If we use squared error loss (as is conventional in NLME modeling), then cross-validation is an estimate of $E\left[\left(Y-\widehat{f}(X)\right)^2\right]$.

A brief explanation of cross-validation is as follows: First, the data is divided into K partitions of roughly equal size. For the kth partition, a model is fit to predict Y based on X using the $K-1$ other partitions of the data. (Note that the kth partition is not used to fit the model.) Then the model is used to predict Y based on X for the data in the kth partition. This process is repeated for $k = 1,2, \ldots, K$, and the K estimates of prediction error are combined. Formally, let $\widehat{f}^{-k}$ be the estimated value of f when the kth partition is removed, and suppose the indices of the observations in the kth partition are contained in K_k. Then the cross-validation estimate of the expected prediction error is equal to

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j\in K_i}\left(y_j-\widehat{f}^{-i}\left(x_j\right)\right)^2$$

Here n denotes the number of observations in the data set. For a more detailed discussion of cross-validation, see Hastie et al (2008).

The above procedure is known as k-fold cross-validation. Leave-one-out cross validation is a special case of k-fold cross-validation where k is equal to the number of observations in the original data set.

2.2 Comparing covariate models

In some situations, a researcher may want to compare models with and without covariate effects, such as a model with an age effect on clearance versus a model without an age effect on clearance. This method is designed to detect differences in models that affect the equations for the parameters.

Consider a data set with subjects $i = 1,2,\ldots,n$. Each subject has observations y_{ij} for $j = 1,2,\ldots,t_i$ (where t_i is the number of time points or discrete values of the independent variable for which there are observations for subject i). The question of interest is whether or not a fixed effect $dPdX$ for a covariate X should be included in an equation for a parameter P, having fixed effect tvP and random effect η_P. The equation for P could have any of the typical forms used in NLME modeling. For example, one could compare a model with a covariate X

$$P=tvP*(X/\text{mean}\,(X))^{dPdX}*\exp\left(\eta_P\right) \quad (1)$$

to a model having no covariate effect

$$P=tvP*\exp\left(\eta_P\right) \quad (2)$$

If a covariate X has an effect on a parameter P, the unexplained error in P (modeled by η_P) when X is left out of the model tends to have higher variance. By including covariate X in the model, we wish to reduce the unexplained error in P, which is represented by η_P. Therefore, metrics involving η_P are useful for determining whether a covariate X is needed. Specifically, one can perform cross-validation to compare the predicted η_P's when X is included or not included in the model. We propose a statistic for determining whether a covariate, X, is needed for explaining variability in a parameter, P, when P is modeled with a random effect η_P. The statistic can be calculated as follows:

For $i = 1$ to n:

1. Remove subject i from the data set.
2. Fit a mixed effects model to the subset of the data with subject i removed.
3. Accept all parameter estimates from this model, and freeze the parameters to those values.
4. Fit the same model to the whole data set, without any major iterations, estimating only the post hoc values of the random effects. (In NONMEM, use the commands MAXITER=0, POSTHOC=Y. In NLME, set NITER to 0.)

5. Square the post hoc eta estimate for the subject that was left out for the parameter of interest

Take the average of the quantity in step 5 over all subjects.

This sequence of steps can also be represented by the equation

$$\mathrm{CrV}_{\eta}=\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\eta}_{P_{i,-i}}\right)^{2} \quad (3)$$

where $\widehat{\eta}_{P_{i,-i}}$ is the post hoc estimate of the random effect for the ith subject for parameter P in a model where the ith subject was removed, and n is the number of subjects. Note that our method leaves out one subject at a time, rather than one observation at a time.

In general, one will favor the model with the minimum value of CrV$_{\eta}$. However, to avoid over-fitting, it is common when applying cross-validation to choose the most parsimonious model (i.e. the model with the fewest covariates) that is within one standard error (SE) of the model with minimum CrV$_{\eta}$ (Hastie et al, 2008). We will follow this convention in all of our subsequent examples. We define SE(CrV$_{\eta}$) as the sample standard deviation of the squared post hoc etas for the subjects left out divided by the square root of the number of subjects. The formula for SE(CrV$_{\eta}$) is given by

$$\mathrm{SE}\left(\mathrm{CrV}_{\eta}\right)=\sqrt{\frac{1}{n\,(n-1)}\sum_{i=1}^{n}(x_i-\overline{x_i})^{2}} \quad (4)$$

where

$$x_i=\widehat{\eta}_{P_{i,-i}}^{2} \quad (5)$$

Alternatively, one may follow the same procedure while removing more than one subject at a time. For example, one may divide the data into k roughly equally-sized partitions, fit a model using the data in $k - 1$ of the partitions, and calculate the post hoc η values for the subjects left out of the model. For data sets with large numbers of subjects, this approach is obviously faster than the “leave-one-out” approach, and it may also reduce the amount of variance in the cross-validation estimates (Hastie et al, 2008). However, this approach may not be practical if the number of subjects is small. We will only consider the leave-one-out method in our subsequent analysis.

2.3 Comparing models with major structural differences

In other situations, a researcher may want to compare models with major structural differences, such as a one compartment model and a two compartment model. This method is designed to detect differences in models that affect the overall shape of the response.

As discussed previously, consider a data set with subjects $i = 1,2,\ldots,n$, where each subject has observations y_{ij} for $j = 1,2,\ldots,t_i$. The statistic can be calculated as follows:

For $i = 1$ to n:

1. Remove subject i from the data set.
2. Fit a mixed effects model to the subset of the data with subject i removed.

3. Accept all parameter estimates from this model, and freeze the parameters to those values.
4. Fit the same model to the whole data set, without any major iterations, estimating only the post hoc values of the random effects. (In NONMEM, use the commands MAXITER=0, POSTHOC=Y. In NLME, set NITER to 0.)
5. Calculate predicted values for subject *i* (the subject that was left out). Note that this estimate uses the post hoc estimate of the random effects for subject *i*.
6. Take the average of the squared individual residuals for the subject that was left out (over all time points or over all values of the independent variable t_i)

Take the average of the quantity in step 6 over all subjects.

This sequence of steps can also be represented by the equation

$$\mathrm{CrV}_y=\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{t_i}\left(y_{ij}-\widehat{y}_{ij,-i}\right)^2}{t_i} \qquad (6)$$

where y_{ij} is the observed value for the *i*th subject at the *j*th time point or independent variable value and $\widehat{y}_{ij,-i}$ is the predicted value for the *i*th subject at the *j*th time point or independent variable value in a model where subject *i* is left out and post hocs are obtained. Once again, note that our method leaves out one subject at a time, rather than one observation at a time.

For purposes of exploration, another statistic that takes into account the weighting of the response can also be calculated:

$$\mathrm{wtCrV}_y=\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{t_i}\mathrm{WTIRES}_{ij,-i}^2}{t_i} \qquad (7)$$

Here $\mathrm{WTIRES}_{ij,-i}$ is the individual weighted residual for subject *i* at time or independent variable value *j* in a model where subject *i* is left out and post hocs are obtained, which is defined to be:

$$\mathrm{WTIRES}_{ij,-i}=\frac{\sqrt{wt_{ij,-i}}\left(y_{ij}-\widehat{y}_{ij,-i}\right)}{\widehat{\sigma}_{-i}} \qquad (8)$$

where $wt_{ij,-i}$ is the weight defined by the residual error model (equal to the squared reciprocal of $\widehat{y}_{ij,-i}$ for constant CV error models or 1 for additive error models), and $\widehat{\sigma}_{-i}^2$ is the estimated residual variance.

As discussed previously, we will follow the convention of choosing the most parsimonious model (defined to the model with the fewest number of compartments) within one standard error (SE) of the model with the minimum CrV_y. The formula for $\mathrm{SE}(\mathrm{CrV}_y)$ is given by

$$\mathrm{SE}\left(\mathrm{CrV}_y\right)=\sqrt{\frac{1}{n\,(n-1)}\sum_{i=1}^{n}(x_i-\overline{x_i})^2} \qquad (9)$$

where

$$x_i = \frac{\sum_{j=1}^{t_i} \left(y_{ij} - \widehat{y}_{ij,-i} \right)^2}{t_i} \quad (10)$$

and the formula for SE(wtCrV$_y$) is calculated similarly, with

$$x_i = \frac{\sum_{j=1}^{t_i} \mathrm{WTIRES}_{ij,-i}^2}{t_i} \quad (11)$$

One may also consider k-fold cross-validation, although we will restrict our attention to leave-one-out for our subsequent analysis.

This method is similar to existing methods for cross-validation on NLME models. However, some applications of cross-validation do not use post hoc estimates of the outcome variable, which is an important difference from our proposed method. Also, we will show why this method should not be used for comparing covariate models.

2.4 Simulated Data Analysis

Five sets of simulated data were generated to evaluate the performance of our proposed cross-validation methods. In each simulation scenario, two models were compared: a sparser "base model" and a less parsimonious "full model." The objective was to determine which of the two possible models was correct using cross-validation.

A brief description of the five simulation scenarios is given in Table 1. For a more detailed description of how the simulated data sets were calculated, see Section S1 in Online Resource 1. For simulation scenarios 1-4, 200 simulated data sets were generated using Pharsight's Trial Simulator software version 2.2.1. (Only 100 simulated data sets were generated for scenario 5.) For each simulated data set, Pharsight's Phoenix NLME (platform version 1.3) was used to fit the appropriate population PK models (both the base model and the full model) using the Lindstrom-Bates method (Lindstrom and Bates, 1990). The η shrinkage of each model was calculated and diagnostics were performed to verify the convergence of each model. To calculate the cross-validation statistics for each simulated data set, subjects were removed from the data set one at a time and the models were recalculated with each subject removed. Post hoc estimates of the random effects (and corresponding predicted values of y) were then calculated for the subject that was excluded from the model. The values of CrV$_\eta$, CrV$_y$, and wtCrV$_y$ were obtained by averaging over each such model. The simulated data sets, batch files, Phoenix mdl files, and other files used to process the output are available from the authors by request.

For each simulated data set, the base model was selected if the value of CrV$_\eta$ for the base model was less than that of the full model. The full model was selected if the value of CrV$_\eta$ was less than that of the base model plus one standard error (using the convention that the more parsimonious model is preferable if its cross-validation error is within one standard error of the cross-validation error of a less parsimonious model). Similar decision rules were used for CrV$_y$ and wtCrV$_y$. The Akaike's insformation criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978) were also calculated for the two models for each simulated data set. The model (base or full) with the smallest AIC/BIC was selected under the two criteria. For each scenario, the performance of cross-validation was compared to the performance of AIC/BIC using a two-sample proportion test.

Note: Consistent with the recommendations of Vonesh and Chinchilli (1997), the BIC was weighted by the number of observations. Although others have suggested that the BIC should be weighted by the number of subjects (Kass and Raftery, 1995), one recent simulation study found that neither choice of weight consistently outperforms the other when applied to mixed models (Gurka, 2006).

2.5 Indomethacin Data Analysis

Pharsights Phoenix NLME (version 1.3) was used to fit models to a previously published indomethacin data set (Kwan et al, 1976). The data consists of six subjects with 11 observations per subject. Each subject was administered a 25 mg dose of indomethacin intravenously at the beginning of the study, and the concentration of indomethacin was measured at 11 time points over an eight-hour period.

The concentrations were plotted versus time for each subject (see Figure 1). Based on the plot, a two compartment IV bolus model with clearance parametrization and a proportional residual error model was fit to this data set. See Section S3.1 in Online Resource 1 for a more detailed description of the model. Individual initial estimates were obtained using the curve stripping method (Gibaldi and Perrier, 1982) with a WinNonlin Classic model. The averages of the individual PK parameters were used as initial estimates for the pop PK model. Random effects were added to the PK parameters for systemic volume and clearance in the form $\theta P * \exp(\eta P)$, where P is the parameter of interest. The Phoenix project file used to fit this model is available from the authors by request.

After fitting the model, a series of diagnostic plots were generated to assess the validity of the model. (See Section S3.2 in Online Resource 1 for details.) The model was then compared to a one compartment model based on both CrV$_y$ and a likelihood ratio test (LRT). First, the model was refit without including any random effects on the PK parameters. (The LRT cannot be used when the random effects are included in this case since the one compartment model forces the removal of some random effects, which implies that these random effects have variances of 0. Thus, comparing the two models would require testing the null hypothesis that a variance is equal to 0, which is on the boundary of the parameter space, rendering the LRT invalid. See Fitzmaurice et al (2011) for details.). The LRT was used to test the null hypothesis of no difference in the predictive accuracy of the two compartment model versus the one compartment model. The value of CrV$_y$ was also calculated for both models. Finally, the value of CrV$_y$ was calculated for a one compartment and two compartment version of the original model (with random effects included). See Section S3.1 in Online Resource 1 for a detailed description of the models that were considered.

2.6 Theophylline Data Analysis

Pharsights Phoenix NLME (version 1.3) was used to fit models to a published theophylline data set (Boeckmann et al, 1992). This theophylline data set consists of twelve subjects with eleven observations per subject. Each subject was administered a dose of theophylline at the beginning of the study ranging between 3.1 mg/kg and 5.86 mg/kg. The concentration of theophylline was measured at 11 time points per subject over a 24 hour period. Each subject's weight was also recorded.

The concentration were plotted versus time for each subject (see Figure 2). Based on the plot, a one compartment extravascular model with clearance parametrization and an additive residual error model was fit to this data set. Random effects were added to the PK parameters for absorption rate, and systemic volume and clearance in the form $\theta\, P \exp(\eta P)$, where P is the parameter of interest. See Section S3.1 in Online Resource 1 for a more detailed description of the model. The Phoenix project file used to fit this model is available from the authors by request.

The LRT and the CrV$_y$ statistic were used to compare a model with a time lag parameter (Tlag) to a model without a Tlag parameter, (with no random effect on the Tlag parameter). Moreover, the covariate plots for the model with Tlag seemed to indicate a body weight effect on *Ka* might be needed (see Figure 3). Thus, the LRT and the CrV$_\eta$ statistic were used to compare a model with the Tlag parameter and body weight effect on *Ka* to the model with the Tlag parameter and no body weight covariate.

3 Results

3.1 Simulation Results

The results of the simulations are summarized in Table 2. The CrV$_\eta$ statistic was correct in 97.0 percent of the 200 cases under scenario 1, whereas AIC was correct in 88.5 percent of cases and BIC was correct in 94.5 percent of cases. It correctly identified the full model under scenario 2 in 92.5 percent of the 200 cases, whereas AIC found the correct model in 98.5 percent of cases and BIC found the correct model in 93 percent of cases. Under scenario 3, CrV$_\eta$ was correct in 93.0 percent of cases, whereas AIC and BIC were correct in 97.5 and 94.0 percent of cases, respectively. Under scenario 4, CrV$_\eta$ was correct in 97.0 percent of cases, whereas AIC and BIC were correct in 71.5 and 64.0 percent of cases, respectively. CrV$_\eta$ was significantly more likely to identify the correct model than AIC in scenario 1 ($p = 0.002$) and it was significantly more likely to identify the correct model than both AIC and BIC in scenario 4 ($p < 0.0001$ in both cases). The performance of CrV$_y$ and wtCrV$_y$ was also evaluated for scenarios 1-4, but it performed poorly in each case with the exception of scenario 1. All four applicable methods (AIC, BIC, CrV$_y$, and wtCrV$_y$) correctly identified the true model under scenario 5 in 100 out of 100 cases.

Some additional information about the distributions of the various test selection statistics are contained in Tables S1, S2, S3, and S4 in Section S1 in Online Resource 1. In general the mean values of AIC and BIC are lower in the true models (compared to the misspecified values) and the mean value of CrV$_\eta$ is significantly lower in the true models. This is not true for CrV$_y$ and wtCrV$_y$ in scenarios 1-4; both statistics tend to be smaller for the base model irrespective of which model is correct (which explains the poor performance of these statistics in scenarios 2-4). It is also note-worthy that an outlying observation generated an extreme value for wtCrV$_y$ for one simulated data set in scenario 3.

Boxplots of the η shrinkage values for both models under scenarios 1-4 are shown in Figure S6 in Online Resource 1. (The η shrinkage values were not calculated for scenario 5 since CrV$_\eta$ was not used in this scenario.) The models converged for all simulated data sets with the exception of two instances of scenario 5 (although some instances of all five simulated scenarios produced models that showed signs of numerical instability).

3.2 Indomethacin Example

The final model appeared to fit well based on the diagnostic plots (see Figures S7, S8, S9 in Online Resource 1). The model coefficients and shrinkage estimates are shown in Tables S5 and S6 in Online Resource 1.

The LRT favored the two compartment model (with no random effects) over the corresponding one compartment model ($p < 0.0001$). The CrV$_y$ statistic was in agreement with the LRT, having a value of 0.1419 (SE 0.03393) for the one compartment model, and 0.0428 (SE 0.01355) for the two compartment model. The CrV$_y$ statistic in the model with random effects also favored the full (two compartment) model over the base (one compartment) model. The value of CrV_y for the full model was 0.01679 (SE 0.004194) and 0.1406 (SE 0.03358) for the base model.

3.3 Theophylline Example

The final model appeared to fit well based on the diagnostic plots (see Figures S10, S11, S12 in Online Resource 1). The model coefficients and shrinkage estimates are shown in Tables S7 and S8 in Online Resource 1.

The LRT favored the Tlag model ($p < 0.0001$). The CrV_y statistic was in agreement with the LRT, having a value of 0.2546 (SE 0.05727) for the model with Tlag, and 0.3927 (SE 0.10001) for the model without Tlag. The LRT had a borderline result ($p = 0.0667$) for comparing the model with a body weight effect on Ka and Tlag to the model without a body weight effect on Ka and Tlag, while the η versus covariate plot (Fig 3) indicated an effect. The CrV_η statistic clearly favored the full model with a Tlag and a weight effect on Ka, having a value of 0.06220 (SE 0.02942) for the full (Tlag and wt) model, and 0.7819 (SE 0.2846) for the base (Tlag) model.

4 Discussion

As noted earlier, cross-validation is not frequently used for comparing NLME models (Brendel et al, 2007). Other methods, such as the LRT, AIC, and BIC are more commonly used. However, each of these alternative approaches have certain drawbacks. All three methods can only be applied to models having the same residual error model. The LRT can only be applied when models are nested and both models have the same random effects. Moreover, there may be an inflated type I error rate associated with the LRTs (Bertrand et al, 2009). Both the AIC and BIC have other shortcomings as well. Specifically, the AIC tends to overfit, meaning that it keeps too many covariates in the model. The BIC, in contrast, tends to underfit (fail to include significant covariates), particularly when the same size is small (Hastie et al, 2008).

Our results show that cross-validation can be used for model selection in NLME modeling and that it can produce significantly better results than these competing methods. The CrV_η statistic identified the correct model at least 92.5% of the time in each of the simulated examples we considered. In contrast, the AIC was significantly more likely to select a covariate for age in our first simulation scenario (even though age had no effect on clearance in the simulated model), and both the AIC and BIC were significantly less likely to detect the effect of hepatic impairment in our fourth simulation scenario.

All four applicable methods (AIC, BIC, CrV_y, and $wtCrV_y$) correctly identified the true (two compartment) in our fifth simulation scenario in 100 out of 100 cases. This finding is of interest because the standard likelihood ratio test cannot be applied when there are random effects in the full model that are not present in the base model.

Both the CrV_y statistic and the LRT favored a two compartment model in the indomethacin example. However, in the theophylline example, the population covariate plots (η's versus covariates) seemed to suggest a weight covariate should be included in the model even though the LRT was not significant at the $p < 0.05$ level. The CrV_η statistic clearly favored the model with the weight covariate. This is consistent with previous studies showing that theophylline distributes poorly into body fat. Hence, the administered mg/kg dose should be calculated on the basis of ideal body weight, (Gal et al, 1978; Rohrbaugh et al, 1982) implying that body weight affects the extent of absorption. This is possible evidence that the LRT cannot always identify covariate effects when they exist and that cross-validation may be able to detect covariate effects in these situations.

Although it may seem reasonable to use the CrV_y or wtCrV_y statistics to determine if a covariate should be included in a model, our simulations suggest that they can give misleading results. Both statistics consistently favored models without a covariate even when a covariate effect existed. The predicted values are just as accurate with and without the covariate effect when the true model has a covariate effect, because the η's can compensate for a missing covariate in a parameter. This suggests that one should use CrV_η rather than CrV_y or wtCrV_y in situations when one wishes to compare different covariate models. This also indicates that it may be misleading to use cross-validation for model validation (as opposed to model selection) if one uses post hoc estimates of the η's when calculating the predicted value of the response on the "left out" portion of the data. One may obtain a low cross-validation error rate even when the model is misspecified.

One possible drawback to using cross-validation for model selection is the fact that it is more computationally intensive than the LRT, AIC, or BIC. Leave-one-out cross-validation was applied in each of the examples in the present study, since each example consisted of relatively small data sets. However, larger data sets may require 1-2 hours (or as many as 10 hours in extreme cases), to fit a single model. If such a data set included hundreds of subjects, leave-one-out cross-validation would clearly be computationally intractible. In such situations, one may reduce the computing time by reducing the number of cross-validation folds. If 10-fold cross-validation is performed, this requires that the model be fitted only 10 times, and the number of folds could be reduced further if needed. Even a complicated model that required ten hours to fit could be evaluated over the course of several days using 5-fold cross-validation. Indeed, these cross-validation methods are no more computationally intensive than bootstrapping, which is commonly used to validate NLME models. The extra computational cost may be worthwhile in situations where it is important that the model is specified correctly.

Another potential issue with cross-validation is the fact that estimation methods for NLME models sometimes fail to converge. Although this was not a major issue in the examples we considered, if the model fails to converge for a significant proportion of the cross-validation folds, it is possible that it will produce inaccurate results. Further research is needed on the effects of lack of convergence on our proposed cross-validation methods.

These methods might be applied more generally with modifications to other types of linear mixed effects models or generalized linear mixed effects models. These methods may be applied without modification to population PK/PD models and sparser data. These are areas for future research. We expect in the sparse data case that the effectiveness of the covariate selection method may be compromised by η shrinkage, which could distort the η size criterion. The covariate selection method introduced in this paper may not produce correct results for parameters with high η shrinkage (greater than 0.3, for example, in a model where the covariate is not included). The random effects for those parameters are typically removed during model development, and hence covariate adjustments may not be needed for those parameters. However, cross-validation produced correct results in our first simulation scenario even though the median shrinkage was approximately 0.3 in the base model (Figure S6 in Online Resource 1). The conditions under which our proposed cross-validation method produces valid results in sparse data sets is another area for future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Eric Bair was partially supported by National Institute of Environmental Health Sciences/National Institutes of Health grant P30ES010126 and National Center for Advancing Translational Sciences/National Institutes of Health grant UL1RR025747. Emily Colby is employed at Pharsight. Bob Leary and Dan Weiner provided valuable input. Tom Colby helped with programming.

References

Akaike H. A new look at the statistical model identification. IEEE Trans Automatic Control. 1974; AC-19:716–723.

Bailey JM, Mora CT, Shafer SL. Pharmacokinetics of propofol in adult patients undergoing coronary revascularization. Anesthesiology. 1996; 84(1):288–97. [PubMed: 8602658]

Bertrand J, Comets E, Laffont CM, Chenel M, Mentre F. Pharmacogenetics and population pharmacokinetics: impact of the design on three tests using the SAEM algorithm. J Pharmacokinet Pharmacodyn. 2009; 36(4):317–339. [PubMed: 19562469]

Boeckmann, A.; Sheiner, L.; Beal, S. NONMEM Users Guide: Part V. University of California; San Francisco: 1992.

Brendel K, Dartois C, Comets E, Lemenuel-Diot A, Laveille C, Tranchand B, Girard P, Laffont C, Mentre F. Are population pharmacokinetic and/or pharmacodynamic models adequately evaluated? A survey of the literature from 2002 to 2004. Clin Pharmacokinet. 2007; 46(3):221–234. [PubMed: 17328581]

Fiset P, Mathers L, Engstrom R, Fitzgerald D, Brand SC, Hsu F, Shafer SL. Pharmacokinetics of computer-controlled alfentanil administration in children undergoing cardiac surgery. Anesthesiology. 1995; 83(5):944–955. [PubMed: 7486179]

Fitzmaurice, G.; Laird, N.; Ware, J. Applied Longitudinal Analysis. 2nd edn.. John Wiley and Sons; Hoboken, NJ: 2011. Wiley Series in Probability and Statistics

Gal P, Jusko WJ, Yurchak AM, Franklin BA. Theophylline disposition in obesity. Clin Pharmacol Ther. 1978; 23(4):438–444. [PubMed: 630791]

Gibaldi, M.; Perrier, D. Pharmacokinetics. 2nd edn.. CRC Press; New York, NY: 1982. Drugs and the Pharmaceutical Sciences Series

Gurka MJ. Selecting the best linear mixed model under REML. The American Statistician. 2006; 60(1):19–26.

Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning: data mining, inference, and prediction. Springer; New York, NY: 2008. Springer Series in Statistics

Hooker A, Ten Tije A, Carducci M, Weber J, Garrett-Mayer E, Gelderblom H, Mcguire W, Verweij J, Karlsson M, Baker S. Population Pharmacokinetic Model for Docetaxel in Patients with Varying Degrees of Liver Function: Incorporating Cytochrome P450 3A Activity Measurements. Clinical Pharmacology and Therapeutics. 2008; 84(1):111–118. [PubMed: 18183036]

Kass RE, Raftery AE. Bayes factors. Journal of the American Statistical Association. 1995; 90(430): 773–795.

Katsube, T.; Khandelwal, A.; Harling, K.; Hooker, AC.; Karlsson, MO. Population Approach Group in Europe 20 (2011). 2011. Evaluation of stepwise covariate model building combined with cross-validation. Abstr 2111, URL www.page-meeting.org/?abstract=2111

Kerbusch T, de Kraker J, Mathôt RA, Beijnen JH. Population pharmacokinetics of ifosfamide and its dechloroethylated and hydroxylated metabolites in children with malignant disease: a sparse sampling approach. Clin Pharmacokinet. 2001; 40(8):615–625. [PubMed: 11523727]

Kwan KC, Breault GO, Umbenhauer ER, McMahon FG, Duggan DE. Kinetics of indomethacin absorption, elimination, and enterohepatic circulation in man. J Pharmacokinet Biopharm. 1976; 4(3):255–280. [PubMed: 978392]

Lindstrom ML, Bates DM. Nonlinear mixed effects models for repeated measures data. Biometrics. 1990; 46(3):673–687. [PubMed: 2242409]

Mulla H, Nabi F, Nichani S, Lawson G, Firmin RK, Upton DR. Population pharmacokinetics of theophylline during paediatric extracorporeal membrane oxygenation. Br J Clin Pharmacol. 2003; 55(1):23–31. [PubMed: 12534637]

Rajagopalan P, Gastonguay MR. Population pharmacokinetics of ciprofloxacin in pediatric patients. J Clin Pharmacol. 2003; 43(7):698–710. [PubMed: 12856383]

Ralph LD, Sandstrom M, Twelves C, Dobbs NA, Thomson AH. Assessment of the validity of a population pharmacokinetic model for epirubicin. Br J Clin Pharmacol. 2006; 62(1):47–55. [PubMed: 16842378]

Ribbing J, Jonsson EN. Cross model validation as a tool for population pharmacokinetic/ pharmacodynamic covariate model building. Population Approach Group in Europe 10 (2001). 2001 Abstr 215, URL www.page-meeting.org/?abstract=215.

Rohrbaugh TM, Danish M, Ragni MC, Yaffe SJ. The effect of obesity on apparent volume of distribution of theophylline. Pediatr Pharmacol. 1982; 2(1):75–83.

Sandler, S. Chemical, Biochemical & Engineering Thermodynamics. 4th edn.. Wiley; Hoboken, NJ: 2006.

Schwarz G. Estimating the dimension of a model. Ann Statist. 1978; 6(2):461–464.

Valodia PN, Seymour MA, McFadyen ML, Miller R, Folb PI. Validation of population pharmacokinetic parameters of phenytoin using the parallel Michaelis-Menten and first-order elimination model. Ther Drug Monit. 2000; 22(3):313–319. [PubMed: 10850399]

Vonesh, E.; Chinchilli, V. Linear and Nonlinear Models for the Analysis of Repeated Measurements. No. v. 1 in Statistics: Textbooks and Monographs. CRC Press; New York, NY: 1997.

Wahlby U, Jonsson EN, Karlsson MO. Assessment of actual significance levels for covariate effects in NONMEM. J Pharmacokinet Pharmacodyn. 2001; 28(3):231–252. [PubMed: 11468939]

Wang Y. Derivation of various NONMEM estimation methods. J Pharmacokinet Pharmacodyn. 2007; 34(5):575–593. [PubMed: 17620001]

Zomorodi K, Donner A, Somma J, Barr J, Sladen R, Ramsay J, Geller E, Shafer SL. Population pharmacokinetics of midazolam administered by target controlled infusion for sedation following coronary artery bypass grafting. Anesthesiology. 1998; 89(6):1418–1429. [PubMed: 9856717]

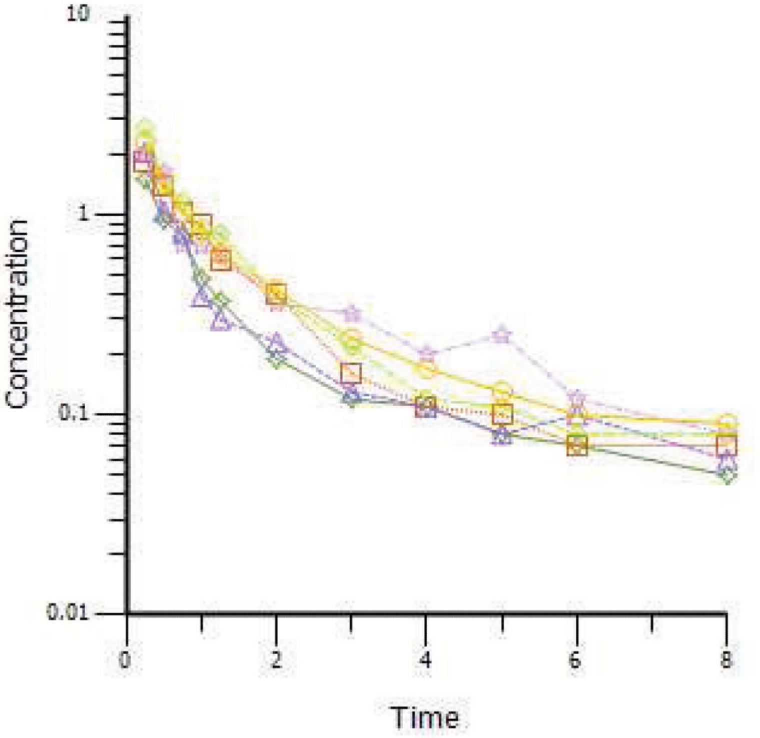

Fig. 1.

Concentration versus time profiles from the indomethacin data set

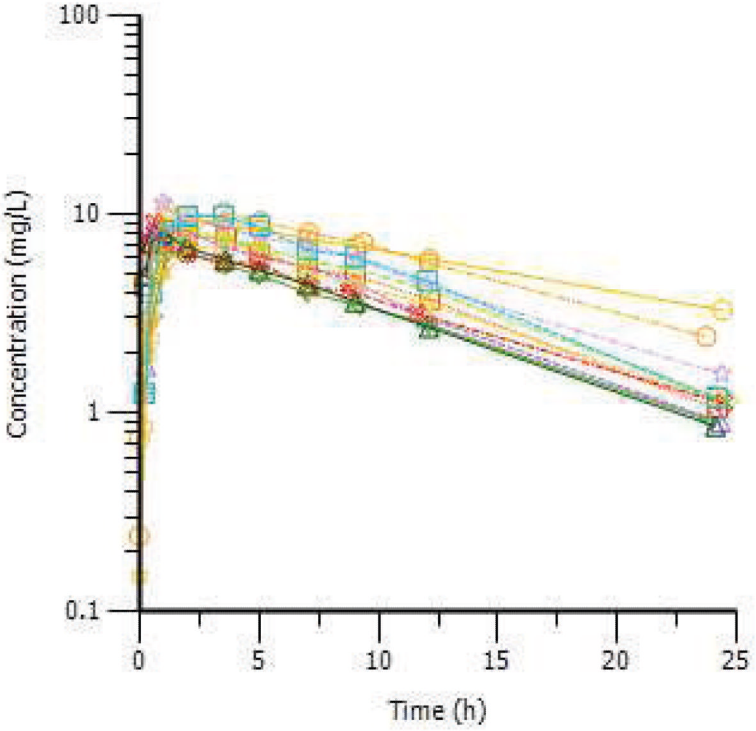

Fig. 2.
Concentration versus time profiles from the theophylline data set

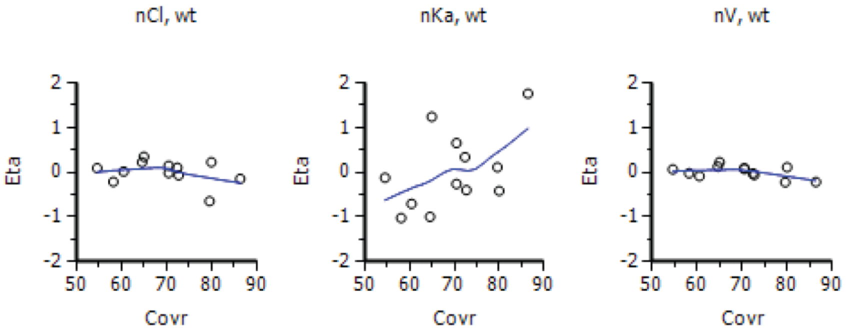

Fig. 3.

η versus covariate plots for the theophylline model with Tlag and no weight effect on Ka

Table 1

Description of the five simulation scenarios

Scen.	True Model	Base Model	Full Model
1	one compartment, no covariate effects	one compartment, no covariate effects	one compartment, age effect on clearance
2	one compartment, age effect on clearance	one compartment, no co-variate effects	one compartment, age effect on clearance
3	two compartments, age effect on clearance	two compartments, no covariate effects	two compartments, age effect on clearance
4	one compartment, body weight effect on volume, body weight, age, gender, and hepatic impairment effects on clearance	one compartment, body weight effect on volume, body weight, age, and gen-der effects on clearance	one compartment, body weight effect on volume, body weight, age, gender, and hepatic impairment effects on clearance
5	two compartments, no covariate effects	one compartment, no covariate effects	two compartments, no covariate effects

Table 2

Proportion of times model comparison methods were correct out of 200 replicates (and associated standard error)

True Model	Comparison	AIC	BIC	$wtCrV_y$	CrV_y	CrV_η
1 Cpt	1 Cpt, Age-Cl	0.885 (0.023)	0.945 (0.016)	0.940 (0.017)	0.965 (0.013)	0.970 (0.012)
1 Cpt, Age-Cl	1 Cpt	0.985 (0.009)	0.930 (0.018)	0.0 (0)	0.0 (0)	0.925 (0.018)
2 Cpt, Age-Cl	2 Cpt	0.975 (0.011)	0.940 (0.017)	0.005 (0.005)	0.010 (0.007)	0.930 (0.018)
1 Cpt, BW-V; BW-Cl, G-Cl, Age-Cl, HI-Cl	1 Cpt, BW-V; BW-Cl, G-Cl, Age-Cl	0.715 (0.032)	0.640 (0.034)	0.015 (0.009)	0.005 (0.005)	0.970 (0.012)
2 Cpt	1 Cpt	1.0 (0) *	1.0 (0) *	1.0 (0) *	1.0 (0) *	N/A

Cpt=Compartment, Age-Cl indicates age effect on clearance, BW=Body Weight, V=Volume, G=Gender, HI=Hepatic Impairment

*Based on 100 replicates