CHAPTER 8
Nonlinear Mixed Effects Models: Practical Issues

In the field of observation, chance favors only the mind that is prepared.

Louis Pasteur (1822–1895), Chemist and the Father of Stereochemistry

Abstract
This chapter provides practical advice in the development of nonlinear mixed effects models. Topics that are discussed include how to choose an estimation method, how to incorporate various covariates in the model (concomitant medications, weight, age, smoking, pregnancy, pharmacogenomics, food, formulation, race, renal function, and laboratory values), and when and how to collect samples for analysis. Missing data and censoring within the context of a repeated measures analysis is also discussed. The chapter concludes with material unique to NONMEM, the leading software for population pharmacokinetic–pharmacodynamic analysis. NONMEM topics that are covered include NONMEM errors and how to resolve them, mu modeling (also called mu-referencing), and the consistency of NONMEM parameter estimates across various platform–compiler combinations.

Introduction
In the last chapter, the theory behind nonlinear mixed effects models was introduced. In this chapter, practical issues related to nonlinear mixed effects modeling will be introduced. Due to space considerations not all topics will be given the coverage they deserve, e.g., handling missing data. What is intended is a broad coverage of problems and issues routinely encountered in actual PopPK analyses. The reader is referred to the original source material and references for further details.

The Data Analysis Plan
Today, good modeling practices dictate that a data analysis plan (DAP) be written prior to any modeling being conducted and prior to any unblinding of the data. The reason being that model credibility is increased to outside reviewers when there is impartiality in the model development process. The DAP essentially provides a blueprint for the analysis. It should provide details about how an analysis will be conducted (notice the emphasis on future tense) and how the results will be reported. Some companies refer to these as Statistical Analysis Plans (SAPs) based on wording in the International Conference on Harmonisation’s (ICH) Guidance on Statistical Principles in Clinical Trials (1998b). Herein these plans will be referred to as DAPs since modeling is not a statistical activity per se.

The 1999 guidance on Population Pharmacokinetics issued by the Food and Drug Administration (FDA) differentiates two types of DAPs, which they refer to as Study Protocols. The first type is an add-on DAP which is seamlessly interwoven into the clinical protocol from which the pharmacokinetic data will be derived. The other type is a stand-alone DAP, which is independent of any clinical protocols, and can “stand-alone” by itself without reference to other protocols. The European Medicines Agency (EMEA) Guideline on Reporting Population Pharmacokinetic Analyses (2007) is written with respect to the stand-alone DAPs which are useful when data from many different studies will be analyzed. In the guidance issued by the EMEA, it is suggested that the essential features of the analysis are included in the clinical protocol, but that the details of the analysis are identified in a stand-alone DAP.

In the stand-alone DAP, the key essential feature is prespecification of the analysis in which the primary analysis variable(s) are defined and methods for dealing with anticipated problems are defined. A DAP differs from the concept of a SAP as defined by the ICH in one important aspect. Modeling is largely an exercise in exploratory data analysis. There are rarely specific hypotheses to be tested. Hence, any PopPK analysis cannot be described in detail as outlined by a SAP under ICH guidelines. Nevertheless, certain elements can be predefined and identified prior to conducting any analysis. But keep in mind that a DAP devoid of detail is essentially meaningless, whereas a DAP that is so detailed will inevitably force the analyst to deviate from the DAP. Hence, there should be a middle ground in the level of detail written into any DAP.

The DAP is usually written by the modeler in conjunction with a physician, a statistician, and, sometimes, someone from data management for analyses of large clinical studies. Some essential elements in a DAP are identified in Table 1. The contents of a DAP can be broken down into signature page, objectives, protocol summary, identification of variables to be analyzed, identification of independent variables, model building procedures, and model validation procedures. The title page and signature page should clearly state the authorship of the DAP, date, and version number. As mentioned, the primary analysis variable and the objectives of the analysis should be clearly defined. In the case of pharmacokinetic data this may be easy to do. In the case of pharmacokinetic–pharmacodynamic modeling, where there may be many possible surrogate markers that may be examined, this may not be easy to do. Particular attention within the DAP should be paid to how the following data types will be dealt with: missing data, censored data, outlier data, time-dependent covariates, and calculation of derived variables. Some DAPs also include detailed “table shells” indicating how all the tables in the report will look when populated with real data. From this, a PopPK report can be easily developed (see Table 2 for a sample table of contents). Once a DAP is finalized, changes can be made prior to unblinding without penalty. However, once the blind is broken and modeling has commenced, all deviations from the DAP should be noted, along with reason and justification for doing so, in the final report.
Table 1
Sample table of contents of a data analysis plan

1. Signature page
2. Abbreviations
3. Objectives and rationale
4. Protocol summaries
   (a) Study design
   (b) Sampling design
5. Subject population(s)
   (a) Description
   (b) Inclusion criteria
   (c) Exclusion criteria
6. Data handling
   (a) Criteria
      - Outliers
      - Missing data
   (b) Subject covariates included in the analysis and rationale
   (c) Subject accounting
   (d) Creation of the NONMEM data file
   (e) Data integrity and computer software
7. Data analysis
   (a) Exploratory data analysis
   (b) Handling missing data
   (c) Population pharmacokinetic model development
      - Assumptions
      - Base model development (if this has been decided)
      - Covariate screening (if screening is to be done)
      - Covariate submodel development
      - Calculated pharmacokinetic parameters
8. Model performance and stability
   (a) Parameter stability
   (b) Sensitivity to model inputs
   (c) Model predictability
   (d) Internal evaluation (bootstrap, influence analysis, etc.)
   (e) External evaluation
9. Pharmacodynamic correlations (if appropriate)
10. Caveats
11. Representative plots and tables
12. Sample report outline (See Table 2)
13. Timelines
14. Miscellaneous considerations
15. References
16. Appendices

Source: European Medicines Agency (EMEA) and Committee for Medicinal Products for Human Use (CHMP) 2007; United States Department of Health and Human Services et al. 1999

*The criteria for model selection and covariate inclusion should be defined in this section*

---

**Choosing an Estimation Method**

With NONMEM (Version 7), the user has a number of estimation algorithms available: first-order (FO) approximation, first-order conditional estimation (FOCE with and without interaction), the hybrid method, Laplacian method, iterative two stage, importance sampling, importance sampling assisted by maximum a posteriori estimation, stochastic approximation by expectation maximization (SAEM), Bayesian estimation via Markov Chain
Monte Carlo (MCMC), and nonparametric estimation. The choice of an estimation method is based on a number of factors, including the type of data, the amount of computation time the user is willing to spend on each run, which is dependent on the complexity of the model, and the degree of nonlinearity of the random effects in the model.

Plainly the most accurate algorithm is desired and at this time, this is probably the Laplacian algorithm. But the Laplacian algorithm is also one of the most computationally expensive and optimization of a single model on current Pentium-speed computers may be prohibitive for routine model development. Balanced against speed is the degree to which the random effects enter into the model nonlinearly. Highly nonlinear models require more accurate algorithms, despite the cost due to increased computation time. In general, models that have nonlinear components, like a
pharmacokinetic model with Michaelis–Menten elimination, population pharmacodynamic models, or models with multiple dosing tend to be more highly nonlinear than single dose pharmacokinetic models, and may require higher numerical accuracy estimation methods. Pharmacodynamic models of ordinal or categorical data should always use the Laplacian method.

First-order approximation used to be the standard method for estimation because FOCE was simply too slow for complex models. With the speed of modern computers and the availability of other estimation options, FO-approximation is no longer used to any significant degree. One situation that FO-approximation can be used is to obtain starting values for more precise estimation algorithms. Obviously, the accuracy of FO-approximation is dependent on the accuracy of the Taylor series approximation to the function and has been shown to provide modestly biased parameter estimates under a variety of different models (Rodman and Silverstein 1990; Sheiner and Beal 1980), particularly when the between-subject and intrasubject variability is high (White et al. 1991). The advantage of FO-approximation is that it is computationally the simplest estimation algorithm and, hence, the fastest. Whether a conditional estimation method will result in parameter estimates much different from estimates obtained using FO-approximation depends on the degree of between-subject variability and the amount of data per subject. When the degree of between-subject variability is small and the model is not highly nonlinear, all the estimation methods tend to produce similar results. Also, the ability of the conditional methods to produce estimates different from FO-approximation decreases when the amount of data per subject decreases.

In general, a good strategy is to use FOCE (with interaction if necessary) for developing the structural and covariate models. With current computer speeds, the algorithm is relatively fast for analytically expressed models. FOCE is still slow when models are expressed as differential equations. Once the final covariate model is identified, other estimation algorithms can be used to compare the stability of the parameter estimates. If FOCE fails, then a stochastic estimation algorithm like SAEM or importance sampling can be tried. Should that fail, a Bayesian method with diffuse priors (or informative priors should that information be available) may be needed.

Systematic comparisons of the different estimation algorithms are few. In general, the newer estimation algorithms, like SAEM and importance sampling, perform as well as FOCE, which is usually used as the reference estimation method. There are situations where a newer estimation algorithm might be preferred to FOCE but these are usually evaluated on a case-by-case basis. For instance, it may be that a base model is more quickly estimated with SAEM than with FOCE, so it may be that SAEM is used for future model development.

At this time, there is limited experience with many of the estimation algorithms in NONMEM 7. Prior to NONMEM 7, some of these algorithms were available in specialized software packages, like the SAEM algorithm in Monolix (http://www.monolix.org) and Bayesian estimation in PKBUGS via the WinBUGS program (http://www.mrc-bsu.cam.ac.uk/bugs/). For a myriad of reasons, penetration of these programs into the modeling community has been slow and many of the publications using these software packages were published by the groups developing them. It will take years before a thorough understanding of which algorithm to use and when to use it is available. For now, most published models still use the FOCE (with interaction) algorithm and the focus of the remainder of this section will be on the performance of the FOCE algorithm.

One reason an estimation method is chosen is that it has a Type I error rate (the probability of rejecting the null hypothesis when true) near nominal values regardless of the experimental design. To examine the Type I error rate using the FO-approximation and FOCE methods, Wahlby, et al. (2001) used Monte Carlo simulation to determine the probability of falsely adding a dichotomous covariate to a model where the covariate was not actually in the model used to generate the data. Pharmacokinetic data were simulated using a default model of a one-compartment model with no covariates and exponential random error. The number of subjects (10–1,000), number of observations per subject (2–19), and size of the residual variance (~10–53%) was varied. Two nested models were used to analyze the data: one without the covariate (reduced model having $p_1$ estimable parameters) and one with a dichotomous covariate introduced on clearance (CL) as a fractional change parameter (full model having $p_2$ estimable parameters). Data sets were simulated and the full and reduced models fit to each data set using FO-approximation, FOCE, and FOCE with interaction (FOCE-I). The change in objective function value ($\Delta$OFV) was calculated for each data set. The proportion of data sets having a $\Delta$OFV greater than 3.84 and 6.63 were determined and deemed the Type I error rate, i.e., the probability of adding another parameter in the model given that the parameter should not be added. Since the OFV produced by NONMEM is proportional to $–2$ times the log-likelihood, the change in OFVs between two nested models should be chi-squared distributed with $p_2–p_1$ degrees of freedom; this is the likelihood ratio test (LRT). In this case, with a single degree of freedom, a $\Delta$OFV of 3.84 and 6.63 should correspond to a significance of 0.05 and 0.01, respectively.

Based on the results of 10,000 simulations with 50 subjects having two observations per subject and residual error of 10%, FO-approximation had a Type I error rate of 0.004, 0.022, and 0.086 for nominal levels of 0.001, 0.01, and 0.05, respectively. FOCE and the Laplacian method were consistently better than FO-approximation but both also had inflated Type I errors. For both FO-approximation and FOCE, regardless of the number of observations per subject, Type I error rates sharply declined when the number of subjects increased from 10 to 50 but thereafter remained relatively constant to 1,000 subjects. In
contrast, Type I error rates with both FO-approximation and FOCE sharply increased when the number of observations per subject increased keeping the number of subjects constant. With 19 observations per subject and 1,000 subjects, the Type I error rate for FO-approximation and FOCE was about 0.47 and 0.37, respectively! Hence, 47 and 37% of simulations, respectively, declared the covariate was an important influence on CL when in fact it was not important. Increasing the residual error to 42% tended to increase the Type I error rate for FO-approximation. FOCE was not studied using the 42% residual error. Interestingly, the inflated Type I error rate seen with FO-approximation and an exponential residual error was not seen when the residual error was homoscedastic (and hence lacking an $\eta\cdot e$ interaction). Also, when the residual variance was modeled using a proportional residual error model, instead of an exponential residual variance model, the Type I error rates decreased with both FO-approximation and FOCE but still remained inflated overall.

Type I error rates with FOCE-I were consistently near nominal values and were unaffected by number of subjects or number of observations per subject. With large residual variability (42%) and two observations per subject, Type I error rates for FOCE-I were higher than nominal, about 0.075 instead of 0.05. But when the number of observations was increased to four, the Type I error rate decreased to the nominal value and remained there as further increases in the number of observations were examined. Also, when the residual variance was modeled using a proportional residual error model, instead of an exponential residual variance model, the Type I error rate decreased. The major conclusion of this analysis was that FOCE-I should be preferred as an estimation method over FO-approximation and FOCE.

Wahlby et al. (2002) later expanded their previous study and used Monte Carlo simulation to examine the Type I error rate under the statistical portion of the model. In all simulations a one-compartment model was used where both between-subject variability and residual variability were modeled using an exponential model. Various combinations were examined: number of observations per subject (2–19), number of subjects (25–1,000), degree of between-subject variability (BSV) and residual variability, and estimation method (FO-approximation, FOCE, and FOCE-I). In their first simulation they compared using the LRT a model having a false covariate influencing interindividual variability in clearance

$$CL = CL_\mu \exp[COV\eta_1 + (1 - COV)\eta_2]$$  \hspace{1cm} (1)

to a model without the covariate, where COV was a dichotomous covariate taking values "0" or "1" denoting which group the $i$th subject was in. The proportion of subjects in each group was 0.5 and both groups were constrained to have the same variance, i.e., $\alpha_{CL,1}^2 = \alpha_{CL,2}^2 = \sigma^2$. In this case, variability was partitioned into two groups. The Type I error rate under FOCE-I remained near the nominal rate of 0.05 as the between-subject variability was increased from 10 to 53% or as the residual error was increased from 10 to 42%. However, the Type I error rate using FO-approximation increased drastically as both between-subject and residual variance increased. For example, the Type I error rate using FO-approximation increased from about 0.06 to 0.15 when the between-subject variability increased from 10 to 52%.

As the number of observations per subject increased from 2 to 19, the Type I error rate using FO-approximation decreased to nominal values, whereas FOCE-I remained near nominal values throughout.

In their second simulation, they examined the Type I error rate when a covariate influenced residual variability, i.e., residual variability was partitioned into two groups. FOCE-I and FO-approximation had a Type I error rate of about 0.075 and 0.11, respectively, with two observations per subject (collected at 1.75 and 7 h after administration) and a residual error of 10%. When residual error increased to 31%, the Type I error rate for FOCE-I decreased to the nominal value, but increased to 0.16 with FO-approximation. With 19 observations per subject, the Type I error rate for both FO-approximation and FOCE-I were unchanged as residual error increased from 10 to 31%, but FOCE-I remained near 0.05.

In their third simulation, they examined the Type I error rate for inclusion of a false covariance term between CL and $V$ in a one-compartment model. FOCE-I Type I error rates were dependent on the number of samples collected per subject (more samples tended to decrease the Type I error rate), degree of residual variability (as residual variability increased so did the Type I error rate), and whether the residual error was defined using an exponential or proportional model (exponential residual error models always produced larger Type I errors than proportional error models). With 100 subjects, two observations per subject (at 1.75 and 7 h after dosing), BSV of ~31%, and residual variability of 10%, the Type I error rate was 0.05 using a nominal expected rate of 0.05 under FOCE-I. The Type I error rate increased to 0.08 when the number of subjects was decreased to 25, but decreased to 0.06 when the number of individuals was increased to 1,000. When the number of observations per subject increased to four (collected at 1, 6, 12, and 18 h) the Type I error rate decreased to the nominal value of 0.05. FO-approximation failed to minimize successfully in most instances with 2 observations per subject, but with 19 observations per subject, even though minimization was often successful, Type I errors were generally elevated.

In their last simulation, the authors examined the Type I error rate when a false variance term was included in the model. In this instance, the boundary issue related to testing diagonal variance terms would be expected to apply and the LRT would not be expected to follow a chi-squared distribution with a single degree of freedom (Stram and Lee 1994). The authors generated concentration–time data with volume of distribution fixed to a constant having no
variability. They then fit a model having $V$ treated as a log-normal random effect. The Type I error rate with FOCE-I was slightly lower than expected when residual error was 10%, regardless of how residual error was modeled. But when residual error was increased and the residual error was modeled using an exponential residual error (which was the same error model used to generate the data), Type I error rates increased significantly. In fact, Type I error rates were greater than 0.85 when residual error was 42% and an exponential residual error model was used! In contrast, Type I error rates remained near their nominal value when a proportional residual error model was used (which was not even how the data were generated). No effect of number of observations per subject, number of subjects, or residual error magnitude was observed when a large number of subjects was combined with a large residual error under FOCE-I, but a significant effect on how the residual error was specified was observed.

In summary, the Type I error rate from using the LRT to test for the inclusion of a covariate in a model was inflated when the data were heteroscedastic and an inappropriate estimation method was used. Type I error rates with FOCE-I were in general near nominal values under most conditions studied and suggest that in most cases FOCE-I should be the estimation method of choice. In contrast, Type I error rates with FO-approximation and FOCE were very dependent on and sensitive to many factors, including number of samples per subject, number of subjects, and how the residual error was defined. The combination of high residual variability with sparse sampling was a particularly disastrous combination using FO-approximation. Type I error rates for testing the inclusion of a variance component in the model were different depending on how the residual error was modeled. Type I error rates were generally higher, and sometimes considerably higher than expected, when an exponential residual error model was used. Slightly lower Type I error rates were observed when a proportional residual error model was used, an effect possibly due to the boundary issues regarding the use of the LRT and the testing of variance components. The authors conclude that the LRT is not reliable when testing the statistical submodel using FO-approximation, but that FOCE-I is reliable when the data per subject are not too sparse and the residuals are normally distributed. Unfortunately, this type of analysis with the new estimation methods has not been conducted.

When using FOCE (with and without interaction) there are two options available: the hybrid method and centering option. The hybrid method is just that, a hybrid between FO-approximation and FOCE. In essence, some elements of the random effects are set equal to zero (the FO-approximation algorithm), while the remaining elements are set equal to the Bayes posterior mode of the random effects (the FOCE algorithm). One advantage to the hybrid method is that it may produce estimates as good as the FOCE method but at a fraction of the computation time. There is little experience in the literature concerning the use of this algorithm.

If after model development is complete and the model still shows systematic bias in the model fit, it may be that the model is misspecified. One assumption of the random effects is that they have zero mean, i.e., they are centered about zero. Whenever one of the conditional estimation algorithms is used within NONMEM, an output of the mean conditional estimate for each random effect, along with a $p$-value testing the null hypothesis that the mean is zero, is produced. Examination of these $p$-values may help identify random effects that are not centered. If many of the random effects are significantly different than zero, a centered model may be more appropriate. Caution should be exercised, however, because in testing many independent random effects “$p$-happens”, i.e., with many independent null hypotheses being tested simultaneously the chances of having at least one significant $p$-value is very high. For example, with a two-compartment model having five random effects the probability of at least one of those conditional means rejecting the null hypothesis may be as high as 23%. When using FOCE or Laplacian estimation with centering, the $p$-values associated with the null hypothesis of the conditional effects may increase. In the worst case, centered models show no difference in their estimates between uncentered models, but may significantly improve a model fit in the best case. Beal and Sheiner (1998) stress that centering should not be used routinely. Like the hybrid method there is little published literature using the centering option. Because there have been no systematic studies comparing these two lesser known methods to the more well-known estimation algorithms, their use should be done with great caution.

In summary, the current state of the art is to use FO-approximation for the first few runs to identify and correct any errors in the data set and control streams. Also, FO-approximation often provides reasonable initial estimates for more computationally intensive algorithms. Thereafter, conditional estimation (usually FOCE-I) is usually used for method development, unless the model is developed using differential equations. In the latter case, even a modestly sized data set with the conditional estimation algorithms may lead to prohibitive run times. Therefore, with these user-written models, other estimation algorithms like SAEM may be needed. After the final model is estimated, other estimation algorithms, like SAEM or importance sampling, can be used to examine the stability of the estimates.

**Issues in the Choice of Covariates**

One issue that a modeler needs to be aware of when testing covariates in a model for statistical significance is when a pharmacokinetic parameter can influence the covariate. Usually it is the other way around – the covariate influences the pharmacokinetic parameter. When the pharmacokinetic parameter influences the covariate a spurious correlation exists between the two that might be misinterpreted as the covariate being predictive of the pharmacokinetic parameter. This issue was highlighted by Ahn et al. (2005) who showed that an artificial correlation is induced between total daily dose (TDD) and clearance.
when analyzing data collected during therapeutic drug monitoring (TDM). This correlation could then be misconstrued as nonlinear pharmacokinetics.

In TDM, patients are dosed and then after a certain period of time their drug concentration in some matrix is measured. If the concentration is below the therapeutic window, the dose is increased. If their concentration is above the therapeutic window, their dose is reduced. And if their concentration is within the therapeutic window, no change in dose is made. In theory, if a person’s concentration is below the therapeutic window, this would imply that their clearance is higher than average. On the other hand, if a person’s concentration is above the therapeutic window then this would imply their clearance is below average. By changing the dose for these patients at the extremes of the distribution for clearance, a correlation is introduced between TDD and clearance, one that in fact has no pharmacokinetic basis.

Using simulation, Ahn et al. (2005) showed that under the TDM design, a clear pattern was seen in weighted residual plots vs. predicted concentrations when clearance was modeled without TDD as a covariate. This trend was also seen in plots of η vs. TDD and in delta plots. In every simulation data set examined, when TDD was introduced as a covariate on clearance the objective function value decreased by at least 42 units; this was significant at \( p < 1 \times 10^{-9} \). The authors conclude that when the pharmacokinetic parameter can influence the value of the covariate, then the covariate is no longer a predictor of the parameter, but a function of it, and that a correlation between the two does not imply causality, a theme that was stressed in the chapter on Linear Regression and Modeling.

Another issue in the choice of covariate in a model is the case where one covariate is correlated with another and acting as a surrogate for the true predictor variable. For example, it may be that age is the true covariate in a model. But because age is highly correlated with weight, in the case where both weight and age are statistically important predictor variables, the effect of weight may be misinterpreted and seen as an effect independent of age, when in fact the weight effect is an artifact. Sometimes the true covariate is unobserved, a so-called latent variable, and the only way to measure its effect is to use a surrogate covariate. But, sometimes the surrogate covariate is not recognized as such.

A case in point was an analysis I did using clofarabine in pediatric patients with solid tumors and hematologic malignancies (Bonate et al. 2004). In that analysis I found a significant relationship between white blood cell (WBC) count and central volume of distribution (V1)

\[
V1 (L) = 115 \left( \frac{\text{Weight}}{40 \text{ kg}} \right) \left( \frac{\text{WBC Count}}{10 \times 10^3 / \mu L} \right)^{0.128}.
\]

When WBC count was removed from the model and using FOCE-I, the LRT increased by 15.2 (\( p < 0.0001 \)) from 2,438.1 to 2,453.3. Hence, WBC appeared to be a very significant predictor for V1. But, WBC count as a covariate for V1 was unusual and raised some flags. I rationalized, nonetheless, that inclusion of WBC count in the model was justified since the site of action for clofarabine was WBCs and since WBC counts decreased from abnormally high values at pretreatment to practically undetectable counts after just a few days of treatment, then the use of this covariate in the model did not seem all that unusual and even had a certain degree of sense to it.

It is recommended in the product label that patients taking clofarabine stay hydrated during therapy since vomiting and diarrhea may occur. It was not until much later when I realized that a more rationale covariate than WBC count was hydration status. A change in plasma volume during the course of therapy might be expected to affect V1 and that V1 on Day 1 might be different than V1 on Day 5. I did not have any predictor for hydration status so when I went back to reanalyzed the data set, I tested a simple model wherein the typical value for V1 (\( \theta \)) was allowed to vary across days

\[
V1 = \begin{cases} 
\theta_1 \left( \frac{\text{Weight}}{40 \text{ kg}} \right) \exp(\eta) & \text{if Day} = 1, \\
\theta_1 \theta_2 \left( \frac{\text{Weight}}{40 \text{ kg}} \right) \exp(\eta) & \text{if Day} = 5,
\end{cases}
\]

where \( \theta_1 \) was the proportional multiplier for Day 5 compared to Day 1. Patients that are dehydrated during therapy might be expected to have a smaller V1 and higher concentrations than on Day 1. This new model resulted in an OFV of 2,435.5, which was lower than the best model using WBC count in the model. Further, \( \theta_2 \) was estimated at 0.86, which was consistent with the hydration status hypothesis.

Hence, what appears to have been happening in the original model was that WBC count, which was changing over time, may have been acting as a surrogate for some time-dependent change in V1 that may have been due to hydration status (or some other unknown latent variable that was changing over time). An alternative view is that (3) is a more flexible, less mechanistic way of modeling (2), which is true. Which model is the right model? There is no right model, but the less mechanistic model using (3) does appear to be slightly better than (2), at least based on OFV. The moral of this story is that a covariate in a model does not necessarily mean that the covariate is in fact the true predictor for some pharmacokinetic parameter, something we may sometimes lose sight of.

**Incorporating Concomitant Medications into the Model**

Identifying important interactions with other drugs is also important for physicians. Zhou (2006) reported that in the 2005 Physician’s Desk Reference electronic library (Version 7.0.308a, 2005.1) 22% of hits using the term “population pharmacokinetic(s)” contained information on drug–drug interactions. Drugs taken in addition to the drug
of interest are called concomitant medications (ConMeds) and are considered extrinsic covariates. How ConMeds are dealt with in a PopPK analysis is unique relative to other covariates in that they are often time-varying and a ubiquitous part of any Phase 3 study. In most Phase 1 studies, ConMeds are prohibited and PopPK analyses based on this data are not confounded by any potential pharmacokinetic interactions with other medications. In contrast, it is almost impossible to stop a patient from taking a ConMed in a Phase 3 study. It is entirely possible that some yet unidentified ConMed will have a clinically significant drug interaction with the drug of interest. A well designed PopPK study can be used to identify such drug–drug interactions. So, dealing with ConMeds becomes an important issue in any PopPK analysis.

One approach to handling ConMeds is to simply classify whether any ConMed is present or absent, regardless of the drug, and then include this binary covariate in an analysis. Interpreting a significant result then is difficult because what exactly does the result mean. That any ConMed affects clearance, for example? Rarely is this approach successful because all the subjects in a study may take a ConMed at some point in time and the entire database will be coded as “Yes.” Also, this approach suffers from low statistical power at detecting anything real. Hence, simple Yes/No coding of ConMed data cannot be recommended. A modification of this approach may be found in oncology in which ConMeds can be coded based on prior therapy. Subjects are coded as no prior therapy (0), minimally pretreated (1), and heavily pretreated (2). While not specific for any one drug, this approach has some appeal.

If one does decide to test for interactions with ConMeds, then it must be decided if testing will be done for specific drugs or groups of drugs. Testing for specific drugs is often done by creating a covariate and assigning it a value of “0” until the ConMed is administered, at which point the value is switched to “1” and remains at “1” until the end of study. The covariate can then be treated as a categorical covariate in a PopPK analysis and the inclusion of the covariate in the model is tested for whether it significantly improves the model’s goodness of fit. Many drug interactions are identified in this manner. This approach, while successful, is not without its problems due to the time of ConMed administration relative to sample collection. More about this will be discussed shortly. A modification to testing for a specific drug is instead of coding as “0” or “1”, group the ConMed into different dose levels, and attempt to quantify the dose–pharmacokinetic interaction effect.

ConMeds can also be classified into groups by either therapeutic area or by mechanism of interaction. For example, many antiepileptic drugs are known to be metabolic inducers. One could then classify whether the ConMed was an antiepileptic or not and use this as a covariate in an analysis. Using such an approach, Yukawa et al. (2002) classified ConMeds into “0” if the ConMed was phenobarbital, phenytoin, or carbamazepine or “1” if none of the above. PopPK analysis of 218 Japanese subjects revealed that these three ConMeds increased haloperidol clearance by 32%. Of course, this approach assumes that these three inducers all have the same time course and effect on the pharmacokinetic parameter that they affect, which may not be the case.

Alternatively, one could group ConMeds into mechanism of action. If a drug is metabolized by cytochrome P450 (CYP), then one such grouping is by CYP isozyme, metabolic pathway or interaction. For example, doxepin is metabolized by CYP 2C19 to desmethyldoxepin. Meyer-Barner et al. (2002) classified concomitant medications into three groups: substrates of CYP 3A4, CYP 2D6, CYP 2C19, CYP 2C9, CYP 2C8, CYP 1A2, CYP 2A6, CYP 2B6, CYP 1A1, and CYP 1B1; inducers of CYP 2E1, CYP 3A4, CYP 3A1, CYP 3A2, CYP 1A2, and CYP 2B6; and inhibitors of CYP 3A4, CYP 1A1, CYP 2C9, CYP 2A6, CYP 2D6, and CYP 2C19. PopPK analysis of 114 patients resulted in the identification of CYP inhibitors as significantly reducing doxepin and desmethyldoxepin clearance by about 15%.

Another confounding factor in the inclusion of ConMed data in an analysis is the time when the ConMed was administered relative to administration of the drug of interest. For example, suppose a ConMed was taken days or weeks prior to the collection of a pharmacokinetic sample. Under the coding scheme just discussed, which is one of the more common coding schemes seen in the literature, one would treat the covariate as “0” until the ConMed is taken, at which time the covariate becomes “1” and remains at “1” thereafter. Hence, at the time the pharmacokinetic sample was taken the ConMed would be coded as “1” or present. This coding may be misleading. If the ConMed was a competitive inhibitor of the drug’s metabolism and ConMed concentrations were far below the inhibition constant at the time the pharmacokinetic sample was collected, then the effect of the ConMed would be negligible. The ConMed would have no influence on the drug’s pharmacokinetics. Clearly, the problem with this approach is that it assumes the ConMed’s effect remains in effect forever. What is needed is a modification to switch the covariate back to “0” at some time after the ConMed is taken, unless the ConMed is taken again.

One such approach is to define time windows around drug administration or blood sample collection and only if the ConMed is taken in the time window is the presence of the ConMed noted. For example, if neutralizing antacids, such as Maalox® or Tums®, are known to affect the absorption of the drug, then a window of ± 2 h around the time of drug administration is created. If the ConMed is taken within ± 2 h of drug administration, the ConMed is coded as “1”; otherwise, the ConMed is coded as “0”. If the ConMed is taken 10 h after drug administration at a time when a blood sample is collected, this sample would be coded as “0” because the drug is already absorbed and the ConMed would have no effect on its kinetics at this point in time. In contrast, suppose the ConMed is phenobarbital,
which is a known rapid and long acting enzyme inducer. One might set up a window of a week after the ConMed was taken, so that any blood sample collected in that time is coded as ‘1’; otherwise, it is coded as ‘0’.

But this is where it again becomes fuzzy. Taking a single dose of phenobarbital does not instantly cause drug induction nor does a single dose affect metabolic induction to the same degree as repeated administration. So how should these facts be taken into account? One might try to recode the variable into $k$-levels denoting the various levels of effect. For example, “1” might be single dose of ConMed, “2” might be multiple dose of ConMed, and “0” might be no ConMed.

Lastly, one other approach to handling ConMeds is to do nothing and not include any ConMeds in an analysis. Many times this is the approach taken by analysts since trying to find important unidentified drug interaction is like trying to find a needle in a haystack. In doing so, however, one accepts that not including an important covariate in a model may lead to larger variance components in the model. Sometimes, in doing an analysis in which ConMeds are not specified as covariates, a subpopulation is identified based on the empirical Bayes estimates (EBEs) of pharmacokinetic parameters or in the concentration–time profiles themselves. In examining those subjects in the subpopulation one may find that they all received the same ConMed. This may cause an analyst to go back, examine all the subjects who received the ConMed, and then include the ConMed in the analysis as a covariate.

But, does lack of a drug interaction in a PopPK analysis mean that the drug interaction does not occur? Of course not. Proving a negative cannot be done. However, by examining the power of the study, one could have some confidence in the conclusion if statistical power was high enough to detect an interaction, but an interaction was not actually detected. More will be discussed on this topic in the section on Experimental Design Issues.

In contrast, detecting a drug interaction raises the question, “is the effect real?” Further studies may be required to confirm such an effect or, in the case of a new drug, the sponsor may choose to include the interaction on the label for safety. An example of the latter is the pharmacokinetic interaction between dolasetron and atenolol identified using PopPK analysis which was later reported on the package insert for dolasetron.

On a practical note, extra effort in the collection of the type of ConMed should be taken for uniformity of spelling or drug name when drug interactions are tested for by specific drug or group. For example, in a Phase 3 study the author participated in, it was found that acetaminophen was coded in the database as acetaminophen, acetominophen, acetamenophen, paracetamol, and Tylenol. Of course, all four were the same thing, but when then database was built, programs like SAS, will not recognize this as such, unless specifically coded for by the programmer. Hence, it is good practice to have someone, preferably a pharmacist, review the ConMeds database for consistency and uniformity of drug names prior to database creation so all ConMeds are accurately accounted for. Second, it should be noted that in reporting analyses involving ConMeds exactly how ConMeds are coded is very important for reproducibility. Many reports in the literature indicate that a ConMed was tested as a covariate, but does not indicate what time window, if any, was used.

**Incorporating Laboratory Tests into the Model**

Laboratory tests, which are surrogate markers for physiological status, are intrinsic covariates that can be altered by other intrinsic and extrinsic factors. Incorporating laboratory test values into a model is done the same as any other continuous type covariate. But, there are a number of issues involved in dealing with laboratory values in PopPK models. First, laboratory values are not time-invariant. Their value may change during the course of a study due to disease progression, other illnesses, the drug itself, the presence of other drugs, dietary changes, regression toward the mean, or by simple random variation, to list just a few causes. Because the covariate may change with time, covariate screening using regression techniques may not be possible. Recall that most regression-based methods regress the empirical Bayes estimate of a parameter against the covariate, a 1:1 relationship. But when the covariate varies over time, which value of the covariate should be regressed against the EBE? One could use the baseline, the mode, or the median, but this approach may be insensitive at detecting a significant relationship. With time-varying covariates, the best approach is to not use regression-based covariate screening and to test the significance of a time-varying covariate directly in the model by performing the LRT on the full and reduced models or use some other measure to see if the goodness of fit of the model is improved when the covariate is added.

Two often encountered situations are where laboratory tests are not collected at the same time as pharmacokinetic samples or when laboratory tests are collected at the same time as the pharmacokinetic samples but are missing for whatever reason. In both cases, the covariate information is missing. Missing data is a fact of life in any population analysis that regulatory authorities recognize. The FDA Guidance on PopPK (1999) states that “missing data will not automatically invalidate an analysis provided a good-faith effort is made to capture the missing data and adequate documentation is made regarding why the data are unavailable.” But unlike age or race, missing laboratory tests are often not completely missing for a subject, rather data are only partially missing. Further, because population studies often have repeated measures in the same individual the usual methods for dealing with missing data, such as those presented in the chapter on Linear Models and Regression, are no longer entirely valid. Handling missing data in a longitudinal study will be dealt with in a later section in this chapter and the reader is referred there for further discussion on this point.
blood count it is not unusual to have 30 potential covariates, 
covariate model. [renal failure and then using this categorical variable in the 
creatinine clearance into normal, mild, moderate, or severe 
Another example along these lines might be classifying 
resulted in a significant improvement in goodness of fit. 
this manner, and including this covariate into the model, 
subjects were classified as normal. Classifying subjects in 
disease were classified as moderate liver disease. Otherwise, 
the normal range and did not meet the criteria for severe liver 
having severe liver disease. Subjects who had values outside 
more than 1.8 mg/dL and a prothrombin time less than 60% as 
Fattinger et al. (1991) classified subjects having a bilirubin 
apopPK analysis of quinidine in 60 subjects with arrhythmia, 
the covariate into severity of disease state. For example, in 
nevirapine clearance than patients with normal values. 
and “0” otherwise. Patients with an aspartate aminotransferase 
laboratory markers of hepatic function into dichotomous 
covariates. Patients were coded as “1” if their laboratory 
Laboratory tests may also show a high degree of 
correlation among each other. For example, aspartate 
aminotransferase (AST) is correlated with alanine 
aminotransferase (ALT) with a correlation coefficient of 
about 0.6 and total protein is correlated with albumin, also 
with a correlation coefficient of about 0.6. Caution needs to 
be exercised when two or more correlated laboratory values 
enter in the covariate model simultaneously because of the 
possible collinearity that may occur (Bonate 1999). Like in 
the linear regression case, inclusion of correlated covariates 
may result in an unstable model leading to inflated standard 
errors and deflated Type I error rate. 
Sometimes, rather than treating the covariate as a 
continuous variable, the covariate will be categorized and 
the categorical variable will be used in the model instead of 
the original value. For example, de Maat et al. (2002) in a 
population analysis of nevirapine categorized baseline 
laboratory markers of hepatic function into dichotomous 
covariates. Patients were coded as “1” if their laboratory 
value was 1.5 times higher than the upper limit of normal and “0” otherwise. Patients with an aspartate aminotransferase 
(AST) 1.5 times higher than normal had a 13% decrease in 
nevirapine clearance than patients with normal values. 
Another example of categorization might be classifying 
the covariate into severity of disease state. For example, in 
a PopPK analysis of quinidine in 60 subjects with arrhythmia, 
Fattinger et al. (1991) classified subjects having a bilirubin 
more than 1.8 mg/dL and a prothrombin time less than 60% as 
having severe liver disease. Subjects who had values outside 
the normal range and did not meet the criteria for severe liver 
disease were classified as moderate liver disease. Otherwise, 
subjects were classified as normal. Classifying subjects in 
this manner, and including this covariate into the model, 
resulted in a significant improvement in goodness of fit. 
Another example along these lines might be classifying 
creatinine clearance into normal, mild, moderate, or severe 
renal failure and then using this categorically variable in the 
covariate model. 
In practice, using a normal chemistry panel and complete 
blood count it is not unusual to have 30 potential covariates, 
everything from sodium ion concentration to alkaline 
phosphatase activity. Early in PopPK analyses it was not 
unusual to screen every single covariate for their impact on 
the model. But a model might end up having a volume of 
distribution as a function of chloride ion concentration or 
clearance that is a function of glucose concentration. 
Physiologically, these covariates are nonsensical. Ideally at 
the end of model development you want a model that is 
interpretable with covariates that make physiologic sense. 
For example, alkaline phosphatase is a marker of hepatic 
function and if alkaline phosphatase ends up as a covariate 
on clearance then the interpretation is easy – hepatic function, 
as measured by alkaline phosphatase, influences clearance. 
But if something unusual, like sodium ion concentration 
influences volume of distribution, then the interpretation is 
difficult and may make no physiologic sense. Hence, the 
current state is to test only those covariates that make 
physiologic sense a priori. So, for example, markers of 
renal or hepatic function (Table 3) will be tested as 
covariates on clearance, while total protein, albumin, and 
α1-acid glycoprotein will be tested as covariates on volume of 
distribution. 
When dealing with time-varying laboratory tests the issue of which value should be modeled often arises. For 
every example, one could use the baseline value, the change from 
baseline value, or the value itself in the model. Obviously 
using the baseline value assumes that the value does not 
change once treatment is initiated, i.e., it removes the time-
varying nature of the value. While sometimes it is believed 
that using change from baseline or percent change from 
baseline will remove any statistical regression toward the 
mean, this is not the case (Bonate 2000). Most analyses use 
the value itself in the analysis and do not bother with either 
baseline or change from baseline. While laboratory tests are 
subject to regression toward the mean, this effect is often 
ignored in an analysis. 
Lastly, one issue that sometimes arises, not during 
model development, but during development of the data set 
used for model building, is when multiple studies are 
combined. First, all medical centers have a clinical chemistry 
laboratory and, as such, each laboratory establishes their 
own “normal reference ranges.” For example, the normal 
range for bilirubin may be 0.4 to 0.7 mg/dL at one lab, but 
0.5 to 0.8 mg/dL at another. Although laboratories attempt 
to establish a high degree of reliability across other labs, the

<table>
<thead>
<tr>
<th>Hepatic function</th>
<th>Renal function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albumin</td>
<td>Creatinine clearance</td>
</tr>
<tr>
<td>Prothrombin time</td>
<td>Serum creatinine</td>
</tr>
<tr>
<td>Bilirubin</td>
<td>Serum urea (BUN)</td>
</tr>
<tr>
<td>Alkaline phosphatase</td>
<td>Urine volume</td>
</tr>
<tr>
<td>Gamma glutamyltransferase (GGT)</td>
<td>Urine urea</td>
</tr>
<tr>
<td>Aspartate aminotransferase (AST)</td>
<td>Urine sodium</td>
</tr>
<tr>
<td>Alanine aminotransferase (AST)</td>
<td>Urine protein</td>
</tr>
<tr>
<td></td>
<td>Urine glucose</td>
</tr>
<tr>
<td></td>
<td>Hematuria</td>
</tr>
</tbody>
</table>
value obtained at one lab might not be the same as at another. When combining data from many study centers there may be laboratory differences. But is this difference of significance and what can be done about it? The long and short of it is, there is nothing that can be done. The analyst must accept the data as is. But, and this especially applies to when studies are combined from the USA and outside the USA, the units that laboratories report data may be different. For example, the USA tends to report concentration data as mass/unit volume, e.g., mg/dL, whereas the rest of the world tends to use moles/unit volume. So when combining data from many studies it is imperative that the same units be used throughout. A slightly more subtle difference may be in reporting scientific notation. For example, a typical white blood cell count may be $6.0 \times 10^9$ cells/L. Some companies store their laboratory data as two variables, value and unit, each within their own study specific database.

In this example, one study database may store the value as “6.0” with units “$10^9$ cells/L,” whereas another study database may store the value as “$6.0 \times 10^9$” and the units as “cells/L.” So, a simple merging of values across databases may result in values ranging from very small to numbers in the thousands. A quick check of the summary statistics (mean, minimum, maximum, and coefficient of variation) for a lab analyte will often reveal whether there are merging errors as the coefficient of variation and range will be very large. Of course, one way to avoid these issues altogether is to use a central laboratory, i.e., all laboratory measurements are done at one central location, which is often done in Phase 3 studies already, so that units and reporting are kept consistent.

In summary, some general guidelines can be recommended in handling laboratory values. Laboratory tests should be identified a priori based on their physiological relevance attached to certain pharmacokinetic parameters and then systematically tested in the model for actual significance. Any data merging should be checked for consistency of units and values by visual inspection or examination of the summary statistics. The data set should also be examined for any missing data. Any missing data should be imputed and clearly identified in the report should be a listing of which values were imputed, the reason for imputation, the method of imputation, and the imputed value. With these guidelines in mind, although some may question the particulars of an analysis, the results will conform to the spirit of regulatory guidelines and be accepted by reviewers.

**Incorporating Weight and Its Variants into the Model**

Weight, an intrinsic covariate, is probably the most frequently identified covariate in a PopPK analysis. All structural model parameters, except absorption-related parameters, may be influenced by a person’s weight. That weight is often identified as an important covariate is not surprising. Since many physiological parameters, such as organ size, organ blood flow, and metabolic rate, are dependent on body weight, then pharmacokinetic parameters that are dependent on blood flow or organ size, should also be a function of body weight. This concept was introduced in previous chapters in regards to allometric scaling.

Weight has many surrogates, such as body surface area (BSA) and body mass index (BMI), any of which may be used in a model. In general, it is best to use a consistent marker for weight throughout an analysis. For example, while modeling clearance as a function of BSA and volume of distribution as a function of actual weight may result in a better model, modeling both clearance and volume of distribution as a function of either weight or BSA alone results in a more harmonious model.

The most frequently used surrogate for weight is BSA and there are many different equations that can be used to calculate a person’s BSA. In the absence of height measurements, Livingston and Lee (2001) developed the following equations for BSA

$$BSA \ (m^2) = 0.1173 \ (Weight \ in \ kg)^{0.6466} \quad (4)$$

for patients weighing more than 10 kg but less than 250 kg and

$$BSA \ (m^2) = 0.1037 \ (Weight \ in \ kg)^{0.6724} \quad (5)$$

for patients weighing less than 10 kg. Sharkey et al. (2001) reported a BSA formula for pediatric and adult patients 10 kg when height was unavailable

$$BSA = 4.688 \times \text{Weight}^{0.8168 \rightarrow 0.0154 \times \text{Ln} (\text{Weight})}$$

and found that the equation deviated by less than 15% from results using other methods that require height and weight as independent variables in every case.

Given a person’s height and weight, more complex equations can be developed, all of which are based on the form

$$BSA = c(Height)^{a_1}(Weight)^{a_2}, \quad (7)$$

where $c$, $a_1$, and $a_2$ are constants. There is no evidence to suggest that these more complex equations using both height and weight are more accurate than Livingston and Lee’s equation. The rationale for such an equation relates to the area of a sphere or cylinder and although the human body is clearly neither of these, the approximation is a useful one. Equation (7) can be linearized by Ln–Ln transformation

$$Ln (BSA) = a_0 + a_1 \times Ln (Height) + a_2 \times Ln (Weight), \quad (8)$$

where $a_0 = Ln(c)$, which will be important later on. Specific equations include DuBois and DuBois (1916)

$$BSA \ (m^2) = 0.007184 (Height \ in \ cm)^{0.725} \ (Weight \ in \ kg)^{0.425}, \quad (9)$$
Table 4

Values of constant terms for various body surface area formulas

<table>
<thead>
<tr>
<th>Equation</th>
<th>$a_0$</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$d_0$</th>
<th>$d_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Du Bois and Du Bois</td>
<td>-4.936</td>
<td>0.725</td>
<td>0.425</td>
<td>-2.41</td>
<td>0.71</td>
</tr>
<tr>
<td>Gehan and George</td>
<td>-3.751</td>
<td>0.442</td>
<td>0.515</td>
<td>-2.28</td>
<td>0.68</td>
</tr>
<tr>
<td>Haycock</td>
<td>-3.719</td>
<td>0.396</td>
<td>0.538</td>
<td>-2.33</td>
<td>0.69</td>
</tr>
<tr>
<td>Mosteller</td>
<td>-4.094</td>
<td>0.500</td>
<td>0.500</td>
<td>-2.35</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Reprinted from Bailey and Briars (1996) with permission

Periodically throughout the literature the validity of the DuBois and DuBois equation, which is the most commonly used predictor in medicinal practice, is challenged since it was based on only nine subjects. Gehan and George (1970) raised this argument and when applied to 401 subjects found that the DuBois and DuBois equation overpredicted BSA by about 15% in about 15% of the cases. Wang et al. (1992) in an examination of 395 subjects, including neonates and pregnant women, compared the precision of 15 BSA prediction formulas and found that the DuBois and DuBois equation underpredicted BSA by about 5%, slightly more in infants. No differences between age or sex were observed in the study and they concluded that the degree of underprediction with the DuBois and DuBois equation is likely to be clinically irrelevant. The DuBois and DuBois equation continues to be the most commonly used predictor of BSA due to its long history of use, but because of its simplicity the Mosteller equation is being increasingly used instead.

Which equation to use in an analysis is for all practical purposes moot since all methods produce equivalent BSA estimates. Bailey and Briars (1996) asked the question “Why do all these formulas produce equivalent results when their constant terms [i.e., $c, a_1$, and $a_2$ in (7)] are so different?” They first analyzed the 401 subjects Gehan and George used to produce their equation and noted that height and weight were highly correlated. Ln-transformed height was a linear function of Ln-transformed weight

$$\ln(\text{Height}) = 3.489 + 0.396 \times \ln(\text{Weight}).$$

When (13) was inserted into (8), Ln-transformed BSA could be determined solely using weight

$$\ln(\text{BSA}) = a_0 + d_1 \times \ln(\text{Weight}),$$

where $a_0 = a_0 + 3.489a_1$ and $d_1 = a_2 + 0.396a_1$. In essence, the relationship linking height and weight to BSA was reduced to a relationship between weight and BSA. Bailey and Briars then computed the estimates of $d_0$ and $d_1$ for the various BSA formula. The results are shown in Table 4. Despite a large difference in constant terms, the values of $d_0$ and $d_1$ were almost exactly the same. The reason then all these equation have such discordant constant terms is parameter instability due to high collinearity between height and weight. When reduced to a more stable model, all these BSA equations are equivalent and any one can be used with no difference in the outcome of an analysis.

Other surrogates for weight are lean body weight (LBW) or fat free mass (FFM), ideal body weight, and BMI. LBW, which has nothing to do with ideal weight, can be calculated in males using

$$\text{LBW} = 1.10(\text{Weight in kg}) - \frac{128(\text{Weight in kg})^2}{[100(\text{Height in m})]^2}.$$ (15)

and in females using

$$\text{LBW} = 1.07(\text{Weight in kg}) - \frac{148(\text{Weight in kg})^2}{[100(\text{Height in m})]^2}.$$ (16)

LBW is the weight of everything in your body except fat. Since fat is relatively inert, it might be expected that LBW might be a better predictor than total body weight in some instances. Examples of where LBW was identified as an important covariate in a PopPK analysis can be found with lithium (Jermain et al. 1991) and carbamazepine (Martin et al. 1991). Use of LBW as a size descriptor has been criticized when the subject is morbidly obese (>30 kg/m$^2$) because such estimates tend to underestimate the subject’s true LBW by overestimating the subject’s percent body fat (Green and Duffull 2002).

Janmahasatian et al. (2005) derived semimechanistically a LBW equation, which they call FFM, using sex, weight, and body mass index

$$\text{FFM} = \begin{cases} \frac{9,270 \times \text{weight}}{6,680 + 216 \times \text{BMI}} & \text{for males,} \\ \frac{9,270 \times \text{weight}}{8,780 + 244 \times \text{BMI}} & \text{for females.} \end{cases}$$ (17)
Since the model development set used to develop (17) included a large proportion of obese subjects, the criticisms ventured against (15) and (16) by Green and Duffull (2002) may be obviated. The equations had good predictive power with a coefficient of determination of 0.93 and a mean error of \(-0.77\) kg.

The second surrogate, ideal body weight (IBW) in kg, can be calculated in males using

\[ \text{IBW} = 52 \, \text{kg} + 1.9 \, \text{kg for every inch over 5 ft} \quad (18) \]

and in females using

\[ \text{IBW} = 49 \, \text{kg} + 1.7 \, \text{kg for every inch over 5 ft} \quad (19) \]

Benezet et al. (1997) presented an interesting use of IBW in a PopPK analysis. They showed that using the mean of IBW and actual body weight resulted in better predictability of carboplatin clearance than actual body weight or IBW alone. In fact, IBW alone underpredicted carboplatin clearance while weight alone overpredicted carboplatin clearance. A rationale for this unusual model was not provided sufficient to say that others had used similar models (Sawyer et al. 1983). While the combined model may have lead to greater predictability than either weight or IBW alone, this unusual model has credibility problems in the absence of a good physiologic rationale. Lastly, BMI is defined as

\[
\text{BMI} \left( \frac{\text{kg}}{\text{m}^2} \right) = \frac{\text{Weight in kg}}{\text{Height in m}^2}. \quad (20)
\]

Normal ranges are from 20 to 26.99 kg/m\(^2\) and apply to both males and females. Rarely is BMI used as a covariate. BSA is more commonly used.

Like laboratory tests, sometimes weight or its surrogates are categorized and the categories used in the model instead. For example, a BMI \(\leq 25\) mg/m\(^2\) would be classified as normal, a BMI between 25 and 29.9 kg/m\(^2\) would be “overweight,” and a BMI > 30 kg/m\(^2\) would be “obese” (National Institutes of Health, 1998). In this manner, the pharmacokinetics between normal and obese subjects can be compared. Examples of the use of obesity in a PopPK analysis are seen with the neuromuscular blockers, doxacurium, and cisatracurium. Schmith et al. (1997) reported in a PopPD analysis of 408 subjects who were administered doxacurium that obesity, defined as actual body weight \(\geq 30\) above IBW, significantly affected the sensitivity of individuals to the drug. In a different analysis of 241 surgical subjects who were administered cisatracurium, obesity (using the same definition) was shown to decrease clearance by 12%, a clinically irrelevant change.

Weight is so important a covariate that in models involving children, infants, or neonates, weight is often built into the model from the very beginning. Weight is treated as an a priori covariate, regardless of whether its inclusion improves the model goodness of fit, for two reasons. First, almost without exception, studies in pediatric populations have shown weight to affect clearance and volume of distribution terms. Second, extrapolations outside the weight range used to develop the model tends to lead to unrealistic predictions.

In pediatric studies, weight is usually modeled as a power function. Whether the exponents to the model are estimated or fixed remains a matter of great debate. There is a strong justification for fixing the exponents to the theoretical values, 0.75 for clearance terms and 1.0 for volume terms, (West et al. 1997), but some argue that allowing the data to “speak for themselves” so-to-speak is a better approach. One advantage of fixing the weight based exponents to their theoretical values is that it allows for further testing of age effects in the model. Allowing both age- and weight-related exponents to be treated as estimable parameters may result in collinearity issues. Fixing the exponents associated with weight removes any collinearity issues in the model.

When the exponential terms for these covariate models are fixed, it must be realized, however, that the equivalent model without weight is not a nested model. For example, suppose clearance was modeled as

\[
\text{CL} = \theta_1 \exp(\eta) \quad (21)
\]

and the fixed exponent model was

\[
\text{CL} = \theta_1 \left( \frac{\text{Weight}}{70 \, \text{kg}} \right)^{0.75} \exp(\eta). \quad (22)
\]

In this case, it should be clear that \(\theta_1\) in (21) is not equal to \(\theta_1\) in (22). Also, there are no degrees of freedom difference between the two equations (they both have 1 estimable parameter) and hence, the two models are not nested. Therefore, the usual chi-squared test for model improvement when weight is included in the model is not valid. In this case, any decrease in the objective function value with the model using (22) compared to the model using (21) is sign of model superiority. However, if the alternative clearance model using weight was

\[
\text{CL} = \theta_1 \left( \frac{\text{Weight}}{70 \, \text{kg}} \right)^{\theta_2} \exp(\eta) \quad (23)
\]

then in this case, the usual chi-squared test does apply because (21) is nested with (23) with 1 degree of freedom.

An illustration of the fixed exponent approach can be found in the PopPK of acetaminophen in premature neonates and infants as presented by Anderson et al. (2002). Data from six previously published pediatric studies were combined into a single analysis of 238 subjects ranging in age from birth to 64 weeks. A one-compartment model was used to characterize the data. Clearance (CL) and volume of distribution (V) were modeled as

\[
\text{CL} = \text{TVCL} \left( \frac{\text{Weight}}{70 \, \text{kg}} \right)^{0.75}, \quad (24)
\]

\[
V = \text{TVV} \left( \frac{\text{Weight}}{70 \, \text{kg}} \right)^{1.0}, \quad (25)
\]
where TVCL and TVV were the typical clearance and volume of distribution values for the population. What was interesting about this analysis was that weight was scaled to a 70 kg adult, even though all patients weighed less than 7 kg. Once weight was included in the model, age could be reliably examined since models with weight excluded tend to “blur” the age effect (Capparelli, personal communication). Other examples of a priori relationships were identified by a survey of published literature. For example, there are many drugs cleared through the kidney, including acyclovir (Tod et al. 2001), ciprofloxacin (Schaefer et al. 1996), and amphotericin B (Mirochnick et al. 1998). Pharmacokinetic data in obese patients is difficult to analyze because the large percent of fat relative to total body weight may make an analysis based on total body weight misleading. They found that no single body size measurement was able to adequately characterize both clearance and volume of distribution equally well in the obese. Most studies used total body weight as a predictor of volume of distribution and appeared to work well for moderate to highly lipophilic drugs. Since clearance does not increase in proportion to total body weight and obese subjects, lean body mass was the best predictor of clearance with no relationship between lipophilicity and preference for the body size descriptor. They suggest that in studies with obese patients, total body weight not be used as a covariate, that possibly lean body mass may be a better covariate.

In summary, weight and its surrogates are important covariates in PopPK analyses. In pediatric studies, weight is often included in the base model as a continuous variable right from the start. In adults, any of the measures of weight can be used as covariates in a model as long as model inclusion improves the goodness of fit and is physiologically plausible. Good modeling practices dictate, however, that whatever variable is used, it is used consistently throughout the model.

**Incorporating Renal Function into the Model**

Next to weight, renal function is probably the most common covariate seen in pharmacokinetic models. As many drugs are cleared through the kidney, it is expected that measures of renal function may be useful in explaining total clearance. For example, enoxaparin is a low-molecular weight heparin analog used in the treatment of deep-vein thrombosis, unstable angina, and non-ST-segment elevation myocardial infarction, whose major adverse events are bleeding-related. Bruno et al. (2003) reported on the population pharmacokinetics–pharmacodynamics of enoxaparin in 448 patients following intravenous and subcutaneous administration. Enoxaparin total clearance (CL, L/h) was found to be described by weight (WT, kg) and creatinine clearance (CrCL, mL/min) by the following function

\[
\text{CL} = 0.733 + 0.00432(\text{WT} - 82) + 0.00338(\text{CrCL} - 87.9). \quad (26)
\]

Based on a logistic regression model using CL as the only covariate, the odds ratio for all hemorrhage and major hemorrhage, the major adverse effects with enoxaparin, was 0.800 (p = 0.0013) and 0.505 (p = 0.0013), respectively. More will be discussed about odds ratios in the chapter on Generalized Linear Mixed Effects Models. Suffice to say that the odds ratio measures the relationship between a covariate and a categorical outcome. In this case, a significant relationship was observed between CL and the adverse event, with decreasing CL leading to an increased risk of experiencing the adverse event.

To understand how this model might be useful in the general population, 1,000 randomly selected male and female subjects from the United States National Health and Nutrition Status Examination Version III (NHANES III) were chosen and their estimated creatinine clearance (eCrCL) was calculated using the Cockcroft–Gault equation (more will be said about this equation shortly) based on the subject’s age, weight, and serum creatinine concentration (SCr). Fig. 1 presents a scatter plot of their eCrCL vs. age. The figure shows that with increasing age, CrCL tends to decrease. Fig. 2 presents a scatter plot of the subject’s predicted enoxaparin total clearance (CL) as a function of eCrCL. With increasing eCrCL, enoxaparin CL increases. Conversely, with increasing age and decreasing eCrCL, it would be expected that enoxaparin CL would decrease, enoxaparin exposure would increase, and the risk of a bleeding event would increase. Fig. 2 also shows a line plot of the probability of all hemorrhagic events and major hemorrhagic events. In both cases, the probability of hemorrhage increases with decreasing CL. Hence, the elderly appear to be at increased risk for an adverse bleeding event.
Nonlinear Mixed Effects Models: Practical Issues

Creatinine Clearance (mL/min)

50 100 150 200

Total Clearance (L/h)

0.0 0.5 1.0 1.5 2.0

Females
Males
LOESS

Fig. 2: Enoxaparin (anti-Factor-Xa) clearance (top plot) as a function of estimated creatinine clearance for the subjects plotted in Fig. 1. Probability of major hemorrhage or all hemorrhage as a function of total enoxaparin clearance (bottom plot).

Based on the results of these analyses which showed that decreased renal clearance leads to an increased risk of an adverse event, the package insert for enoxaparin, which is marketed as Lovenox®, reads:

Because exposure of enoxaparin sodium is significantly increased in patients with severe renal impairment (creatinine clearance <30 mL/min), a dosage adjustment is recommended for therapeutic and prophylactic dosage ranges. No dosage adjustment is recommended in patients with moderate (creatinine clearance 30–50 mL/min) and mild (creatinine clearance 50–80 mL/min) renal impairment.

Interestingly, weight was also a covariate in the submodel for CL but no dose adjustment appears to be needed for obesity. This example was meant to illustrate how a covariate can be used to aid in making dose decisions for patients, and in particular how estimates of renal function can be used to make dose modifications.

The kidneys and liver are the primary organs of elimination from the body. Since total clearance is the sum of renal, nonrenal, and other routes of elimination, it makes sense to include a marker of renal function as a covariate in the submodel for a drug’s clearance, like in (26). But what markers of renal function should be used in the covariate submodel? Renal clearance (CLR) is the sum of glomerular filtration, tubular secretion, tubular reabsorption (which negatively affects CLR), and renal metabolism (Bonate et al. 1998). The latter is often ignored in theoretical analyses of renal function. Hence,

\[
\text{CLR} = \text{CL}_{\text{Filtration}} + \text{CL}_{\text{Secretion}} - \text{CL}_{\text{Reabsorption}}. \tag{27}
\]

Filtration clearance is the volume of fluid that is cleared of drug by the glomerulus per unit time and is equal to \(fu \times \text{GFR}\), where \(fu\) is the unbound fraction of drug and GFR is the glomerular filtration rate. Secretion clearance, which is transporter-mediated, is equal to

\[
\text{CL}_{\text{Secretion}} = \frac{T_{\text{max}}}{T_m + C_p}, \tag{28}
\]

where \(T_{\text{max}}\) is the maximal rate of secretion from the blood to the tubules, \(T_m\) is the concentration in plasma that results in 50% of the maximal rate of secretion, and \(C_p\) is the plasma concentration. Reabsorption clearance is the fraction of filtration clearance and secretion clearance that is reabsorbed back into the blood.

Thus one could use surrogates for each of these clearance terms based on a known reference and then relate a drug’s clearance to a function of the surrogates. So if a drug were primarily eliminated by filtration, the drug’s clearance could be written as

\[
\text{CL} = \theta_1 (fu \times \text{GFR}) + \text{CL}_{\text{nonrenal}}. \tag{29}
\]

In practice, however, the unbound fraction of the drug is ignored and GFR used alone. GFR is generally regarded as the best index of renal function since its decline precedes renal failure and is an indicator of chronic kidney disease.

The gold standard of GFR is the urinary clearance of a substance that is freely cleared by the kidney and has no metabolism, like \(^{51}\)Cr-EDTA, inulin, iothalamate, or cystatin-C. These markers are difficult to use in practice since their use requires exogenous intravenous administration. After administration, blood or urine samples are collected, assayed for the substance, and then total systemic clearance or urinary clearance, respectively, is calculated and equated to GFR. Even though their routine use is difficult, this is not to say that they are not used. Viberg et al. (2006) reported that cystatin-C clearance was a better predictor of cefuroxime clearance than serum creatinine and other routine measures of renal function.

Because of the difficulties just mentioned, researchers have searched for endogenous markers of pure glomerular filtration. One early marker thought to meet the criteria for a marker of GFR (not protein bound, freely filtered at the glomerulus, not subject to reabsorption or secretion, physiologically inert, and easily measurable) was urea, the primary product of protein metabolism. While blood urea nitrogen (BUN) is freely diffusible across the glomerulus,
its concentration can be altered by a high-protein diet, hydration status, and gastrointestinal bleeding and is not sensitive enough to be very useful since more than two-thirds of renal function must be lost before notable changes in BUN occur.

Another proposed marker is serum creatinine since creatinine is freely diffusible across the glomerulus, is not metabolized, and is relatively constant in the plasma. However, creatinine suffers from the same problem as urea, namely diet. While serum creatinine concentration could be used as a marker of renal function, it too is insensitive to alterations in renal function since more than half of renal function must be lost before changes in serum creatinine occur.

A more common measure is urinary creatinine clearance (CrCL), which is often characterized through the use of 24 h urine collections, and measures the volume of plasma or serum cleared of creatinine per unit time. CrCL is not an ideal surrogate for GFR since CrCL has a small contribution from tubular secretion, thereby overestimating GFR by as much as 20%, can be influenced by drugs that might present in plasma, and is influenced by muscle mass. There are also practical problem such as inadequate urine collection intervals or missed urinary collection.

Because of the problems associated with actual measurement of CrCL, many researchers have sought to develop predictive models of CrCL based on patient characteristics and clinical chemistry values. The most common predictive model is the Cockcroft–Gault equation in which the estimated CrCL (eCrCL in mL/min) is a function of sex, age (in years), weight (in kg), and serum creatinine concentration (Scr, mg/dL)

\[
eCrCL = \frac{(140 - \text{age}) \times \text{Weight}}{72 \times \text{Scr}} \times (0.85 \text{ if female}) \tag{30}
\]

or in the International System (SI) of units

\[
eCrCL = k \frac{(140 - \text{age}) \times \text{Weight}}{\text{Scr} \times \text{Scr in } \mu\text{mol/L}} \tag{31}
\]

where \( k \) is 1.04 for women and 1.23 for men. Generally eCrCL is capped at 200 mL/min because of physiological constraints.

The predictors in the Cockcroft–Gault equation have a high degree of face validity. Age is included in the model since younger adults have higher GFR than older adults having the same creatinine clearance, an effect most likely due to younger adults having higher muscle mass, and subsequent creatinine production, than older adults. The same rationale applies to having a correction factor for females since males have higher muscle mass than females. And lastly there is an inverse relationship between serum creatinine and creatinine clearance with high serum creatinine concentrations corresponding to low creatinine clearance.

The Cockcroft–Gault model was derived from 249 heterogeneous male veterans (no females were included) aged 18 to 92 years old and is the preferred method by many regulatory agencies although that may change with the advent of the Modification of Diet in Renal Disease (MDRD) equation, which is now recommended by the National Kidney Foundation and American Nephrology Society as the preferred method for estimating GFR.

The MDRD equation as originally proposed, which was developed in 1,628 patients with a mean age of 50.6 years in consisting of 983 males and 645 females, does not use weight, but instead uses race, serum creatinine, age, and sex and is given by

\[
eCrCL = 186 \left( \frac{\text{Scr}}{88.4} \right)^{-1.154} \times \text{Age}^{-0.203} \times (0.742 \text{ if female})(1.21 \text{ if Black}) \tag{32}
\]

and for SI units is given by

\[
eCrCL = 32788 \left( \frac{\text{Scr}}{88.4} \right)^{-1.154} \times \text{Age}^{-0.203} \times (0.742 \text{ if female})(1.21 \text{ if Black}). \tag{33}
\]

In both cases, eCrCL is in units of mL/min/1.73 m², which differs from the Cockcroft–Gault equation which is given in units of mL/min. The MDR equation is highly predictive having a coefficient of determination of 89.2%.

Caution must be exercised when using the MDRD using data prior to 2009. Serum creatinine has mostly been measured analytically using the Jaffe reaction, in which creatinine turns orange when mixed with picric acid under alkaline conditions. The problem is that this colorimetric assay is not completely specific for creatinine and can measure other “proteins” as well. Over time, various versions of the original reaction have been proposed to decrease the bias associated with noncreatinine chromagens. The originally proposed MDRD equation was based on the Jaffe reaction and used a kinetic alkaline picrate assay, whereas the Cockcroft–Gault equation used an alkaline picrate reaction with deproteinization. Fortunately, the overestimation of serum creatinine and subsequent underestimation of eCrCL tended to be offset by the degree of tubular secretion so that the two biases canceled each other out and eCrCL actually appeared unbiased for GFR.

However, in the early 2000s it was realized that the current assays for creatinine needed to be standardized across laboratories to a greater extent than was currently done. A national standardization program was implemented in most clinical chemistry laboratories in the USA in the late 2000s using “gold-standard” serum creatinine reference materials such that serum creatinine concentrations will be lower than before and eCrCL estimates will be higher than before. The problem this presents is that when analyzing data across studies in drug development programs that span the decade of the 2000s it may be that some studies were done using the standardized assay and some with the older method. Since the reaction actually used by a laboratory is not collected in clinical studies it is impossible to know which study used which assay (Stevens and Levey 2009). Hence, the situation may present where there is a differential eCrCL effect across studies. For studies after 2010, this issue should be a nonissue since most labs will be using the standardized assay.
The new standardized assay is based on the enzymatic reaction of creatinine results in a slightly modified form of the original equation

\[
eCrCL = 175 \left(\frac{SCr}{88.4}\right)^{-1.154} \text{Age}^{-0.203} \times (0.742 \text{ if female})(1.21 \text{ if Black}),
\]

where the intercept has been modified to account for the bias with the previous analytical method. It should be noted that the MDRD equation has been shown to be less accurate in subjects without kidney disease and in patients about to undergo kidney donation. The MDRD equation has not been validated in pediatric populations, pregnant women, the very elderly (>85 years), in the presence of some concomitant medications, and in some ethnic groups such as Hispanics. Further, an unreported problem with the MDRD equation is its reliance on race as a covariate and all the inaccuracies that are involved in categorizing patients on their race, e.g., how should a mixed race patient be categorized? Nevertheless, studies comparing the Cockroft-Gault equation to the MDRD equation, the MDRD performs as well as or better than the Cockroft-Gault equation in most instances, particularly in older and obese patients.

The Cockroft–Gault equation and the MDRD equation should not be used in pediatric populations. The National Kidney Foundation’s Kidney Disease Outcomes Quality Initiative (K-DOQI) clinical practice guidelines for chronic kidney disease in children (Hogg et al. 2003) recommends the Schwartz equation (Schwartz et al. 1976)

\[
eCrCL (\text{mL/min/1.73 m}^2) = \frac{55 \times \text{Height in cm}}{\text{SCr in mg/dL}},
\]

which was derived in 186 children between the ages of 6 months to 20 years, or the Counahan et al. (1976) equation

\[
GFR (\text{mL/min/1.73 m}^2) = \frac{43 \times \text{Height in cm}}{\text{SCR in mg/dL}},
\]

which was derived in 108 subjects, 103 of which were children between the ages of two months and 14 years, and 5 adults. Both equations are of the same functional form with the only difference being the value of the constant, 55 in the Schwartz equation and 43 in the Counahan equation. This difference in the constant is apparently due to differences in how serum creatinine was measured in the two studies. Both equations tend to overestimate eCrCL with decreasing GFR.

Pierrat et al. (2003) compared the Cockroft–Gault equation, Schwartz equation, and MDRD equation in the 198 children (with two kidneys, single kidney, and transplanted) and 116 adults (single kidney and transplanted) using inulin clearance as the reference. They recommend that in children 12 years and older, the Cockroft–Gault equation be used but that its value not be interpreted as creatinine clearance but GFR standardized to body surface area. Using the Cockroft–Gault equation, 95% of predictions are within ±40 mL/min/1.73 m². In children 12 years and younger no formula was satisfactory at predicting GFR.

Recently, a new equation has been developed that is reported to be even more accurate than the MDRD equation. This equation, called the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation, was developed in 8,254 participants in the model development data set and 3,896 participants in the model validation data set, with GFR values ranging from 2 to 190 mL/min/1.73 m², and was developed using standardized creatinine measurements. The CKD-EPI equation is considerably more complex than previous equations as it involves changing parameters at a knot in the serum concentration continuum

\[
eCrCL = 141 \times \min \left(\frac{SCr}{k}, \frac{1}{1.08}\right)^{0.993} \times \max \left(\frac{SCr}{k}, 1\right)^{1.209} \times 0.993^{\text{AGE}} \times \left(1.108 \text{ if female}\right) \times \left(1.159 \text{ if Black}\right),
\]

where \( k \) equals 0.9 for males and 0.7 for females and a equals −0.411 for males and −0.329 for females, \( \min(\cdot) \) indicates the minimum of either \( SCr/k \) or 1, and \( \max(\cdot) \) indicates the maximum of either \( SCr/k \) or 1. The CKD-EPI equation performed better than the MDRD equation did, especially at higher GFRs, with a median difference of 2.5 mL/min/1.73 m² for the CKD-EPI equation compared to 5.5 mL/min/1.73 m² for the MDRD equation. Studies in pharmacokinetics literature are few using this equation in time tell utility of this new equation.

Practically, the significance of eCrCL as a covariate in a population model is tested by placing eCrCL as a linear predictor on clearance. There is no rationale for testing eCrCL as a covariate on the other model parameters like volume of distribution, thus limiting the number of models that need to be tested. As an example, Jullien et al. (2008) studied the pharmacokinetics of tenofovir in patients with HIV and examined the role of renal function in tenofovir clearance using SCr, eCrCL, weight divided by Scr as covariates. Scr decreased the OFV by 33, eCrCL decreased the OFV by 44, and body weight divided by SCR decreased the OFV by 55. Further, interindividual variability decreased by 16, 29, and 39%, for Scr, eCrCL, and body weight divided by SCR, respectively. The final model for clearance was

\[
CL (L/h) = 90.9 \left(\frac{\text{Weight}}{\text{SCr}}ight)^{0.83} \times \left(\frac{L}{T}\right)^{0.77},
\]

where \( L \) was 0.86 if tenofovir was combined with lopinavir/ritonavir and \( T \) was 2.3 if tubulopathy was observed.

It should be noted that a problem arises when analyzing data from adults and children simultaneously when renal function is an important predictor of the drug’s pharmacokinetics. For example, if eCrCL is measured in adults using the MDRD equation, and eCrCL is measured in children using the Schwartz criteria, there will be a discontinuity at the point where the pediatric population converts to an adult population. Unfortunately there is no solution to this problem.
Incorporating a Food Effect into the Model

Food, an extrinsic covariate, has been shown to affect the pharmacokinetics of many orally administered drugs, both in a positive and negative manner, through its effect on the absorption process (Singh 1999). In most Phase I studies, drug administration is done in the fasted state, whereas drug administration may be in the either fasted or fed state in Phase 2 or 3. Hence, testing of a “food effect” is commonly done in studies of a mixed nature where the drug can be given either with or without food. As might be surmised, the simplest method for testing of food effect is to create a new covariate (e.g., FOOD) that applies to each dosing interval where samples are collected such that the covariate is coded as either “1” if the drug is administered with food or “0” if the drug is not given with food. The covariate can then be applied to either the absorption rate constant, lag-time, or relative bioavailability (F1), e.g.,

\[ \text{F1} = \theta_1 (1 + \theta_2 \text{FOOD}), \]

where \( \theta_1 \) is the baseline relative bioavailability to the dosing compartment in the absence of food (\( \text{FOOD} = 0 \)) and \( 1 + \theta_2 \) is the proportional multiplier in the presence of food (\( \text{FOOD} = 1 \)).

Of course, like the issue of drug interactions, the timing of when the meal is given is an issue. For solid foodstuff, the half-life of stomach emptying is about 90–120 min (West 1985). So, one option is to code the covariate as “1” if the drug is taken within 90–120 min of a meal; otherwise, the covariate is coded as “0.” For example, suppose for the same subject that pharmacokinetic samples are collected on Day 1 and 5 of a 5-day dosing regimen and that food was given 1 h prior to drug administration on Day 5, but not on Day 1. Then a new covariate called FOOD can be created which is equal to “1” on Day 1 but “0” on Day 5. In this manner, Van Wart et al. (2004) showed that garenoxacin, a new fluoroquinolone antibiotic, does not exhibit a food effect. Coding is more difficult if food is taken after drug administration, but whatever coding scheme is applied must be consistently applied and specified in the data analysis plan.

The type of meal that is administered may also affect the pharmacokinetics of a drug. For example, a light snack might have no effect but a heavy, fatty meal might. Hence, an alternative coding scheme, instead of “0/1,” would be to use a categorical scheme where “1” might be a light meal or breakfast, “2” might be lunch, and “3” might be dinner. Another method might be to break the meals down into a series of dichotomous variables, one for each meal. Bonate (2003) used such a scheme to quantify the food effect after placebo administration on QTc interval prolongation. He further treated the food effect as not an “either/or” proposition but as an exponential function

\[ \text{QTc} = \begin{cases} 
\theta_1 + \theta_2 \exp(-\theta_3 t) & \text{if breakfast}, \\
\theta_1 + \theta_3 \exp(-\theta_4 t) & \text{if lunch}, \\
\theta_1 + \theta_4 \exp(-\theta_5 t) & \text{if dinner},
\end{cases} \]

In this manner, the food effect increases to its maximal immediately after a meal but then declines in an exponential manner. At baseline (\( \theta_i \)), QTc intervals were 389 ms but increased by 10.6, 12.5, and 14.7 ms after breakfast (\( \theta_2 \)), lunch (\( \theta_3 \)), and dinner (\( \theta_4 \)). Further, the rate of decline in food effect was estimated at 0.4 ms/h (\( \theta_5 \)). Hence, the heavier the meal the larger the increase in QTc intervals.

Incorporating Patient Age into the Model

Age, an intrinsic covariate, is commonly identified in population analyses as being important since many physiological processes change with age. For example, Gilmore et al. (1992) found that propranolol intrinsic clearance in elderly subjects greater than 62 years old was 30% lower than in subjects 25 to 33 years old and that the elimination half-life was two to three times longer in the elderly than in the young. The mechanism for difference has been attributed to reduced hepatic blood flow in the elderly. Clearly, understanding age-related changes in pharmacokinetics and pharmacodynamics is important from a therapeutic perspective. For recent reviews on the mechanisms around such age-related changes, the reader is referred to Mangoni and Jackson (2004) who focus on changes in the elderly and Loebstein et al. (2000) who focus on age-related changes in children.

Often, age is identified as important when the age range in the database is large. Data pooled from Phase 1 studies, where the typical age is from 18 years to the mid-40s, has insufficient signal to identify age as a covariate since the range is so narrow. But, when data from Phase 1 is pooled with data from elderly subjects or pediatric patients, then the signal is usually large enough to detect an age effect, if it is important.

Age can be treated as either a continuous variable in the model, wherein it is typically centered around the mean or median, or it can be treated as a categorical variable. When age is treated as a continuous variable in adults, its value is typically an integer. So if someone was 32 years and 11 months old, they would be treated as simply 32 years old. This truncation is not problematic for adults. But truncation error becomes significant as patients get younger and younger so that age is usually not truncated with pediatric patients. Indeed, for the very young, a few months may have a significant impact on the pharmacokinetics of a drug.

Categorization of age can be done using regulatory documents as guidelines. The (1994) ICH Guidance on studies in pediatric patients (E11) breaks down pediatric patients into four groups:

1. Preterm newborn infants
2. Term newborn infants (0–27 days)
3. Infant and toddlers (28 days to 23 months)
4. Children (2–11 years)
5. Adolescents (11 to <16–18 years, depending on region)

Subjects older than 18 years are expected to have similar pharmacokinetics–pharmacodynamics as adults. It should be noted that the FDA uses slightly different categories for
pediatric patients (United States Department of Health and Human Services et al. 1998). The ICH Guidance on Studies in Support of Special Populations: Geriatrics (E7) FDA Guidance on Drugs Likely to be Used in the Elderly (1994) defines an elderly subject as 65 years or older. Although it may appear reasonable to model pediatric or elderly data using dummy variables based on these categories, this approach may result in loss of information following categorization. Nevertheless, it is common to see this categorization approach used in comparing elderly to young subjects, but it is uncommon in modeling pediatric data. More often, age is treated as a continuous variable after controlling for weight in pediatric studies.

In adults, age is defined based on postpartum age, i.e., time at birth to the date of first dose administration or date of randomization in a study. Normal gestation is defined as 40 weeks with term infants being born after 38–42 weeks gestation. Gestational age is defined as the length of time from conception to delivery. When modeling data from neonates, term newborn infants, and even up to infants and toddlers, postpartum age may not accurately represent the physiological status of the patient if the child was born prematurely (less than 36 weeks gestation) or preterm (36–37 weeks gestation). For this reason, postconceptual age, which is the gestational age plus postpartum age, may be a more accurate reflection of physiological age than postpartum age. Anderson et al. (2002) showed that postconceptual age was an important covariate in predicting acetaminophen apparent volume of distribution and apparent oral clearance in premature neonates and infants. Also, gestational age itself may be a useful covariate in an analysis since premature newborns are more likely to have physiological problems, like respiratory distress, than term newborns. For example, Grimsley and Thomson (1999) found that gestational age, coded as “0” for greater than 35 weeks and “1” for less than 35 weeks, was an important predictor of vancomycin clearance in neonates. Lastly, as an aside, age is typically confounded with weight in a pediatric population and it is difficult to separate the effect of each on a drug’s pharmacokinetics. In the modeling of a drug administered to a pediatric population, weight is frequently built into the model a priori (see section on incorporating weight into the model), although this practice is not universally accepted. Once weight is built into the model, age can then usually be added to the model and tested for model improvement.

Incorporating Formulation Effects and Route of Administration into the Model

Drug formulation, an extrinsic covariate, plays a major role in the rate and extent of absorption and, hence, in a drug’s pharmacokinetic profile. In the development of new orally administered drugs it is not uncommon for the formulation to change during the development process. For example, the first-time-in-man study may administer the drug as a solution, then as a capsule in later Phase I studies, and then as a tablet during Phase 2. Perhaps, later, some small change is made in the design of the tablet, like a change in the excipients, which is then the final formulation used throughout Phase 3. The bottom line is that rarely is the marketed formulation used throughout the development process. Yet in a population pharmacokinetic analysis using data across all phases of development, these differences in the formulation must be accounted for. For example, the rate and extent of absorption of the solution formulation may be different than a capsule which may be different than a tablet.

An important concept in the absorption modeling of any extravascular administered drug is that the absorption process is independent of distribution and elimination. Hence, in accounting for these different formulations all that needs to be accounted for is the absorption process because distribution and elimination is usually the same regardless of how the drug is absorbed. Hence, different formulations are usually accorded their own absorption model through a series of IF-THEN statements with one formulation being the reference formulation. Thus, a solution and capsule may be modeled as

\[
\begin{align*}
\theta_4 &= \theta_1 \exp(\eta_1) \\
F_1 &= 1 \\
\text{ALAG} &= \theta_2 \exp(\eta_2) \\
k_3 &= \theta_3 \exp(\eta_3) \\
F_1 &= \theta_4 \exp(\eta_4) \\
\text{ALAG} &= \theta_5 \exp(\eta_5)
\end{align*}
\]

where both the solution and capsule are modeled using a first-order absorption model with different rate constants (defined as \(k_a\), \(\theta_1\) and \(\theta_3\), and lag-times (defined as \(\text{ALAG}\)), \(\theta_2\) and \(\theta_5\). Under this model, one formulation must serve as the reference formulation having a bioavailability (defined as \(F_1\), where the “1” indicates that bioavailability references compartment number 1) of 1, even though the true bioavailability may not equal one. Hence, \(\theta_i\) measures the bioavailability relative to the reference formulation. In this instance, all estimable parameters were treated as random effects, but this is not a requirement. Some, none, or all of the parameters in the model could have been treated as random effects; it depends on the data. BSV associated with absorption is rarely estimable with sparse data and most absorption parameters are treated as fixed effects. It should be noted that other distribution and clearance-related parameters for the drug are then shared between the different formulations. In other words, if the drug follows a two-compartment model, then clearance, intercompartmental clearance, central volume, and peripheral volume are the same for both routes of distribution.

When a drug is given by two different routes of administration, the same trick is used to model the absorption process. In this case, however, sometimes it is necessary to specify that dosing is into different compartments. For example, suppose the drug follows one-compartment kinetics and the drug can be administered by the intravenous or oral route of administration. In this case, dosing with the intravenous route is into the central compartment, whereas dosing is into a dosing compartment...
after oral administration. Which compartment the drug is administered into must be correctly defined for correct identification of model parameters.

**Incorporating Smoking Status into the Model**

Smoking has been shown to affect the metabolism of many drugs, such as warfarin and theophylline, through induction of liver enzymes which leads to increased clearance (Miller 1989). As such, smoking status has been added as a covariate in many population pharmacokinetic analyses usually in a simple dichotomous form, e.g.,

\[
CL = \theta_1 (1 + \theta_2 \times \text{SMOK}) \exp(\eta),
\]

where \(\theta_1\) is the population mean for nonsmokers, \(\theta_2\) is the proportional multiplier for smokers, and \(\text{SMOK}\) is defined as “0” for nonsmokers and “1” for smokers. For example, Ng et al. (2009) reported in a population analysis of 197 patients with schizophrenia that clozapine clearance was increased by 33% in smokers compared to nonsmokers leading to reduced clozapine exposure in smokers. Bigos et al. (2008) reported in a population analysis of 523 schizophrenia patients that olanzapine clearance was increased 55% in smokers. While smoking has been identified as an important covariate in some studies, the literature is more replete with examples where smoking was not an important covariate. This may be due to difficulty with which to classify smoking status.

In the clozapine report by Ng et al., smoking status was defined as “current smoker” vs. “nonsmoker.” If it only were that simple. Smoking research tends to consider many different variables. Leffondre et al. (2002) report that the following all must be considered as part of any research into the effect of smoking on some outcome:

- Smoking status: never smoked, current smoker, and ex-smoker (at least 1 year prior to study enrollment)
- Time since cessation of smoking: 1 day to 2 years, 2–5 years, 5–10 years, 10–15 years, and >15 years
- Intensity of smoking (cigarettes/day)
- Duration of smoking (years)
- Cigarette-years (cigarettes/day × duration of smoking)
- Age at initiation of smoking (years)

Leffondre et al. comment on the difficulty of using so many metrics in an analysis of smoking data, particularly with regards to multicollinearity. As a result, Hoffmann and Bergmann (2003) derived a “Comprehensive Smoking Index” (CSI) as a single aggregate to reflect smoking exposure which incorporates total duration of smoking, time since smoking cessation, smoking intensity, and selected interactions between these variables. Specifically, the CSI was defined as

\[
\text{CSI} = \left(1 - 0.5^{\text{dur} / \tau}\right) \left(0.5^{\text{tsc} / \tau}\right) \text{int},
\]

where \(\text{dur}\) is the duration of smoking, \(\text{tsc}\) is the time since stopping smoking, \(\text{int}\) is the number of cigarettes per day, and \(\tau\) is the decline in smoking effect over time. Notice that the CSI consists of three components. The first component accounts for how long a person has smoked, the second accounts for if a patient has stopped smoking, and the last component accounts for how much they smoked(d). \(\text{Dur}, \text{tsc},\) and \(\text{int}\) are all known quantities but \(\tau\) depends on the effect being measured. Dietrich and Hoffman (2004) reported \(\tau\) to be 1.5 years in an analysis of chronic periodontitis data. The CSI has not yet been used in pharmacokinetics or drug metabolism so it is difficult to report appropriate values of \(\tau\) in this regard. The following table presents various combinations of duration, recency, and intensity and the corresponding CSI value:

<table>
<thead>
<tr>
<th>Intensity</th>
<th>Duration</th>
<th>Recency</th>
<th>(\tau)</th>
<th>CSI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Current Smokers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
<td>1.5</td>
<td>4.5</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>0</td>
<td>1.5</td>
<td>9.0</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>0</td>
<td>1.5</td>
<td>18.0</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>0</td>
<td>1.5</td>
<td>5.0</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0</td>
<td>1.5</td>
<td>9.9</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>0</td>
<td>1.5</td>
<td>19.8</td>
</tr>
<tr>
<td><strong>Past Smokers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>2</td>
<td>1.5</td>
<td>3.9</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>2</td>
<td>1.5</td>
<td>7.9</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1.5</td>
<td>1.0</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>5</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>1.5</td>
<td>0.1</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>10</td>
<td>1.5</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Leffondre et al. (2006) has suggested that (43) may have some residual confounding issues, which may occur when there is confounding with unmeasured, latent variables or measurement error in the independent variables. As such, they suggest that the model

\[
\text{CSI} = \left(1 - 0.5^{\text{dur} / \tau}\right) \left(0.5^{\text{tsc} / \tau}\right) \text{int} \left(\text{int} + 1\right),
\]

where \(\text{dur}\) is \(\max(0, \text{tsc} - \delta, 0)\) and \(\text{tsc}\) is \(\max(0, \text{dur} + \text{tsc} - \delta, 0)\) is more appropriate. Equation (44) introduces a new estimable parameter \(\delta\) which can be thought of as a parameter to account for a lag-time in effect.

Equation (43) can be used to model clearance in a population model

\[
\text{CL} = \theta_1 \left(1 - 0.5^{\text{dur} / \tau}\right) \left(0.5^{\text{tsc} / \tau}\right) \text{int} \exp(\eta).
\]

Equation (44) can be expanded into a similar form for CL. In both instances \(\theta\) is the set of estimable parameters. No one has used the CSI in a population analysis and its utility remains to be seen. On face value, its validity seems high. Besides lack of use, another disadvantage to using the CSI is that it requires more information than is usually collected during a clinical trial and will require some planning prior to the start of the study.

**Incorporating Race into the Model**

Decades of research have shown race to be an important determinant of drug effect for some, but not all, drugs (Johnson 1997). One example is the lower rate of efficacy of
β-blockers in Blacks than Whites in the treatment of hypertension. Examples of race as an important covariate in the PopPK literature are few. One example is etanercept in the treatment of rheumatoid arthritis. Lee et al. (2003) reported that apparent clearance was 38% higher in nonwhites compared to whites. Interestingly, however, this race effect was not observed when etanercept pharmacokinetics were studied in patients with psoriasis (Nesterov et al. 2004). Although it would be easy to conclude that the difference is due to some combination of pharmacokinetic, possibly due to genetic differences in drug metabolizing capability, and pharmacodynamic alterations, the reality is that these differences may also be due to other nonspecific factors like diet, geography, and differences in health care or weight. Nevertheless, it is important to test race as a covariate in a population analysis as it may help to reduce BSV, despite its potential nonspecific causality.

Possibly to avoid the social ramifications of using the word “race,” often in the medical literature one sees “race/ethnicity” or simply “ethnicity,” as if these were measuring the same thing. They do not. Ethnicity relates to the cultural associations a person belongs to. For example, a person from the Dominican Republic may be classified as “Hispanic or Latino” on the US Census but their ethnicity may relate more closely with Blacks. According to the International Conference on Harmonisation (ICH) Guideline on Ethnic Factors in the Acceptability of Foreign Clinical Data (1998a), ethnicity has a broader meaning than race as it encompasses cultural as well as genetic constructs; race falsely implies a biological context. How race and ethnicity are used in combination will be discussed shortly.

Although as scientists we use “race” as if it were a scientific term, race is a social construct—not a biological one. Prior to 1989, a child was classified on their birth certificate in the USA as White only if both parents were White (LaVeist 1994). A child was classified as Black if the father was Black (regardless of the mother’s race) or if the mother was Black (regardless of the father, unless the father was Japanese in which case the child was classified as Japanese). Confusing? Imagine how it is for a child of mixed race ancestry. In contrast, at the same time in Japan, a child was declared Japanese only if the father was Japanese, regardless of the race of the mother. Current standards in the USA still vary from state to state with most states offering a variety of options. In the 2000, US Census more than 126 possible racial and ethnic categories were available. Hence, how a person is racially classified is dependent on the country of origin and the current sociopolitical constructs at the time, which can make for problems when pooling racial data from across the globe.

That race is not a biological construct rests on a number of considerations (Williams 1997). First, the concept of race predates biological attempts at classification; people have always been classified by their skin color and other external features. Second, the phenotypic expression used to classify an individual, like skin color, does not correlate with genetic variability. It is often stated that there is more variability within races than between races. In other words, two individuals from any two races will have about as much genetic variability as any two individuals from within the same race. While considerable biological variation exists across groups, there is just as much variation within groups. Nevertheless, race is still wrongly viewed as a valid biological construct as it is routinely used for differential medical diagnoses, treatment, and care. And because of this false belief in its biological validity, race has a notorious history for discrimination.

Although the reader may intuitively understand what race is and how to measure it, as it is used ubiquitously throughout the medical literature, there are clearly measurement and conceptual issues that must first be resolved before it can be used in an analysis. One problem with race is that many people believe they can define it, usually based on a person’s skin color. It is crazy to believe that a person’s skin color is a predictor for complex physiological differences, yet this is what we do as scientists. Further, racial classification is a categorical decision based on a continuous scale. Skin color varies from light to dark. Where is the break-point and who decides? Clearly, race is a surrogate for some other factor: socioeconomic, dietary, genetic, etc. There must be a better way to measure such differences.

External assessment of racial assignment is called physionomy but is often confounded by ethnocentricity, which means that assessment tends to be in reference to the assigner’s own ethnicity. This raises many issues, the least of which is intrarater and interrater consistency and misclassification rate. Boehmer et al. (2002) found a high rate of discordance between internal and external assessment of race in 12,444 patients surveyed for race in a Veterans Administration study of dental outpatients. A total of 14% of Hispanics, 1.5% of Whites, and 5% of Blacks were incorrectly racially classified according to the patient. If such classification is required to be done by external assessment then perhaps a more objective, quantitative measure is needed. Klag et al. (1991) used a light meter to measure the relative lightness or darkness of a person’s skin and then used that information to determine the effect of skin color on that person’s blood pressure. This approach has not yet been explored in clinical trials to any significant extent.

One approach that has been explored is the use of genetic markers to define race. Given that there are three billion bases in human DNA with 99.9% of these bases exactly the same across individuals, this means that each individual has more than three million unique base pairs. There are too many unique base pairs per individual for classification using all unique base pairs. Hence, the base pairs are grouped into a few manageable categories or

---

1 Current estimates from analysis of variance show that differences among major groups account for at most 5% of the total variability in genetic variation with within-population differences among individuals accounting for the remainder, upwards of 95% (Rosenberg et al. 2002).
clusters which are then used to define a particular racial group. Wilson et al. (2001) used microsatellite markers to infer genetic clusters from eight racial populations. Using four genetic clusters they found substantial misclassification; 21% of Afro-Caribbeans were clustered with West Eurasians, while only 24% of Ethiopians were classified with Afro-Caribbeans. A total of 62% of Ethiopians were placed in the same cluster as Jews, Norwegians, and Armenians. Hence, genetic analyses appear to be inadequate to categorizing individuals into groups of sufficient homogeneity to be useful (Risch et al. 2002). Further, “commonly used ethnic labels (such as Black, Caucasian, and Asian) are insufficient and inaccurate descriptions of human genetic variation” (Wilson et al. 2001).

The most common and now recommended approach for identifying race is self-identification. A recent editorial in the Journal of the American Medical Association (JAMA) states that “individuals should self-designate race to ensure that the designation most closely matches what they believe reflects their personal and cultural background” (Winker 2004). The International Committee of Medical Journal Editors, of which JAMA belongs, now requires that authors define how race was measured and to justify their relevance in the analysis. Questions may be either open ended, e.g., race: ______, or closed-ended, e.g., Race: (1) White, (2) Black, etc., provided the coding process is transparent. If the questions are closed-ended then the categories should be presented, whether categories were combined, and if so, how they were combined.

This approach, while simple, also appears to be practical. One biotechnology company, NitroMed (Lexington, MA) has recently submitted a New Drug Application with the FDA for BiDil®, a combination tablet containing hydralazine plus isosorbide dinitrate, for the treatment of heart failure in African-Americans. Non-African-Americans were specifically excluded from the clinical trial – a first for clinical drug trials (Franciosa et al. 2002). African-Americans were accepted into the trial based on self-designation. In their pivotal trial, a 43% improvement in survival was demonstrated, along with a 33% reduction in first hospitalizations and better quality of life. In 2005, the FDA Cardiac and Renal Advisory Committee recommended approval of the drug, despite concerns by the FDA that the effect observed was based on “old-fashioned way of determining race which relies on one’s perception of one’s race.” Nevertheless, the medical reviewer at the FDA concluded that “finally a drug is probably able to efficiently control blood pressure in [African-Americans] and prevent the consequences of both hypertension and [heart failure].” Shortly thereafter, the FDA formally approved BiDil for the treatment of “treatment of heart failure as an adjunct to standard therapy in self-identified black patients to improve survival, to prolong time to hospitalization for heart failure, and to improve patient-reported functional status.” Interestingly, other races are not listed as special populations in the package insert.

Unfortunately for multiracial studies, blind use of common categories like White, Black, Asian, Hispanic, or Other is too limiting. In 1998, the FDA issued the demographic rule to reflect that some individuals may respond differently to drugs and to make such analyses looking for racial subgroup differences consistent across regulatory submissions (United States Department of Health and Human Services et al. 2003). In their guidance the FDA was required to implement the initiatives published by the Office of Management and Budget (OMB), guidelines that were required of all federal agencies. The following guidelines are suggested:

1. Use a two-question self-reporting format requesting race and ethnicity with ethnicity preceding race to allow for multiracial identities. When self-reporting cannot be done then such information should be provided by a first-degree relative or other knowledgeable source.

2. For ethnicity, the following minimum classes should be offered: (a) Hispanic or Latino and (b) Not Hispanic or Latino.

3. For race the following minimum choices should be made: American Indian or Alaska Native; Asian; Black or African American; Native Hawaiian or Other Pacific Islander; and White.

For example, Hispanic or Latino refers to any person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin, regardless of race. Still, there might be some confusion with their implementation (for example, should a person from Portugal be classified as Spanish or White?) so internal consistency should be stressed. The guideline does allow for more flexibility in the collection of race and ethnicity, e.g., White can be subgrouped to reflect European White or North African White. However, any use of more expansive categories must somehow reflect the original five racial categories suggested. The reader is referred to the guidelines for more details. As an aside, under the new guidelines issued by the OMB, in the 2000 United States Census, the first-time individuals were allowed to identify themselves with more than one race, over 7 million people said they did in fact belong to more than one race, thus highlighting the need for diversity of racial categorization.

The OMB guidelines were not based in science and this fact is clearly stated; the categories were sociopolitical constructs created for expediency. In a review article on racial categorization, Risch et al. (2002) state that the human race is best categorized into the following groups: Africans, Caucasians, Pacific Islanders, East Asians, and Native Americans. Notice that Hispanics are not included in the list because Hispanics are a mix of Native American, Caucasian, and African/African American with varying regional proportions. For example, Southwest American Hispanics are a mix of Native American (39%), Caucasian (58%), and African (3%). East Coast Hispanics have a greater proportion of African admixture. Thus, depending
on geographic location, Hispanics could coaggregate more similar to Caucasians, Native Americans, or African-Americans.

Current guidelines by the FDA and National Institutes of Health mandate that safety data be analyzed by race. Such blanket requirements are unfortunate because they reinforce the impression that race is responsible for possible outcome disparities without regard to cause and effect. Hence, these analyses become “check-box” analyses. Within the context of population analyses, the testing of race as a covariate is not required, thus leaving it to the modeler’s discretion. If race is tested as a covariate, its rationale should be justified.

The actual testing of race as a covariate is done using dummy variables. For instance, suppose subjects in a study were either White (coded as “0”) or Black (coded as “1”) and clearance (CL) was being modeled. Then using the dummy variable approach with a proportional change model, the effect of race on CL could be modeled as

\[ CL = \theta_1 (1 + \text{RACE} \times \theta_2) \exp(\eta), \]  

(46)

where \( \theta_1 \) represents the CL in the White participants and \( 1 + \theta_2 \) is the CL multiplier between Whites and Blacks. So, if \( \theta_2 \) were 0.28 then CL in Blacks would be 28% higher than CL in Whites. It could also be assumed that race, not only affects the population mean, but the variability in CL such that the model could now become

\[ CL = \theta_1 (1 + \text{RACE} \times \theta_2) \exp(\eta_2 \times \text{RACE} + (1 - \text{RACE}) \eta_1), \]  

(47)

where now \( \theta_1 \) and \( \theta_2 \) are defined as before but \( \eta_1 \) and \( \eta_2 \) define the deviation from the population mean for White and Black subjects, respectively. For more than two categories, IF-THEN statements can be used to define the racial covariate model.

It has been stated that usually at least 80% of participants in clinical trials in the USA are White (Holden 2003). Hence, non-White subjects in other particular groups may have few observations such that it would be almost impossible to detect any significant pharmacokinetic–pharmacodynamic differences among the groups. So as a matter of practice, those subjects with few observations are grouped into some other group. Immediately, however, the question arises as to which group should another group be placed into. Suppose there were only a handful of Asians in a clinical trial involving hundreds of participants, most of which are White or Black. Should the Asian participants be placed into the White group or Black group? Clearly, this practice has the potential to obscure any real differences among groups.

But does identifying a racial difference imply a biological difference? Not necessarily. Identifying race as a covariate does not imply that there is some underlying genetic or biological difference between groups but that in the absence of other information subjects within one racial category have some characteristics in common allowing them to differ from others in another racial group.

How race will be used in clinical practice raises some interesting problems in their own right. For instance, if a racial difference is identified, its use in clinical practice requires a physician to be aware of the racial difference, make an external assessment of the patient’s race, and then identify whether the patient falls into an at-risk category. Further, a patient may not agree with the physician’s racial assessment and hence may refuse the differential treatment guidelines. These problems are fortunately outside the modelers realm.

It is important to remember that the finding of race as a covariate carries more political baggage than other covariates, like age or weight, because of its long history as a means to discriminate. It has been argued that emphasizing biological differences will lead to greater discrimination and differences in health care. Others argue that studying racial differences will lead to better health care and improved treatment for patient subgroups. Still, others argue against using race at all in an analysis since it has no proven value for the individual patient (Schwartz 2001). Although race is fraught with problems in its measurement, its nonbiologic basis, and its implementation may be a useful surrogate for some other variable, like diet, geography, or environment, and to ignore it is akin to sticking one’s head in the sand. A compromise must be achieved that maintains the dignity of those classes affected by the racial model.

**Incorporating Pregnancy into the Model**

With pregnancy a number of physiological changes occur in the mother-to-be: changes in gastric pH and protein binding, increased extracellular fluid, enzyme induction, and increased blood flow, just to name a few (Anderson 2005; Reynolds and Knott 1989). These changes can result in changes in protein binding, clearance, volume of distribution, intercompartmental clearance, and drug absorption. There are few reports in the literature using the population approach, but given the difficulty with sampling and studying pregnant women in general, the population approach is the preferred method of analysis.

Most often, testing whether pregnancy is an important covariate in a model is done by treating pregnancy as a binary categorical variable ("0" if the patient is not pregnant and “1” if the patient is pregnant) and then using a proportional multiplier model, e.g.,

\[ CL = \theta_0 (1 + \theta_1 \times \text{PREGNANCY}). \]  

(48)

An example is the study by Karunajeewa et al. (2009) who showed a 33% increase in sulfadoxine clearance in pregnant women compared to nonpregnant women. Capparelli et al. (2008) showed no difference in nevirapine pharmacokinetics between pregnant and nonpregnant women with human immunodeficiency virus (HIV).

However, this approach fails to account for the fact that pregnancy is a dynamic process. A pregnant woman in her first trimester might not have the same physiological changes as in the last trimester. A binary covariate treats a woman who has just gotten pregnant similarly as a women just about to give birth. In fact, a woman in her first trimester may be more like a nonpregnant woman than a pregnant woman in her last trimester. Hence, it may be
reasonable to set some time cutoff where only if the woman is pregnant beyond a certain number of weeks are they considered “pregnant” for the sake of the model. For example, Bouillon-Pichault et al. (2009) used 15 weeks gestational age (which is the age of the fetus or newborn infant usually defined as the number of weeks from the time of the woman’s last menstrual period or from 2 weeks before conception) as the cutoff for defining pregnancy in a population analysis of the pharmacokinetics of lopinavir in pregnant women with HIV; this value was chosen empirically by testing various cutoff values and using the one with the lowest OFV. In this manner, lopinavir clearance was shown to be 39% higher in pregnant women having a gestational age of 15 weeks or higher compared to nonpregnant women and women of less than 15 weeks gestational age.

Other types of covariates related to pregnancy may prove useful. For example, if a woman’s weight prior to pregnancy is known then the difference between their current weight and baseline prepregnancy weight, which represents the added weight due to the fetus, may be a useful covariate. Degree of edema may also be useful as a covariate. Muller et al. (2008) used a four-level scale for degree of edema: 0, no edema to 3, above the knee. They found that amoxicillin’s central volume of distribution increased with increasing level of edema in pregnant women. Similarly, the trimester of a woman’s pregnancy may be a more useful covariate than simply a binary yes pregnant vs. not pregnant covariate.

Related to pregnancy are the changes in pharmacokinetics after giving birth. The few studies examining these changes have used a binary covariate “antepartum vs. postpartum” to test for differences. Capparelli et al. (2008) showed no difference in nevirapine pharmacokinetics in women antepartum and postpartum.

As of yet there are few reports in the literature on the pharmacokinetics during pregnancy. It has not been that long ago that children were called “therapeutic orphans” because of the lack of drug-related research in children. With recent legislation that is changing. But pregnant orphans” and hopefully in the near future that may change as well.

**Incorporating Pharmacogenomics into the Model**

Despite the improvement in goodness of fit when significant covariates are incorporated into a pharmacokinetic model, in reality, the addition of most covariates reduce the unexplained variability in a model by very little. There are very few “silver bullet” covariates where their addition to the model has such an effect that residual variability is reduced by orders of magnitude or all the between-subject variability is explained by the covariate. One example of this might be a drug that is not metabolized and excreted entirely by the kidneys, such as with Org31540/SR90107A, a pentasaccharide, which is eliminated virtually 100% by the kidney as parent drug (Faaji et al. 1998). Creatinine clearance accounts for 90% of the variability in total clearance. Beyond this, most covariates fail to account for the majority of variance, even in the best examples of PopPK analysis.

One potential “silver bullet” that is only now being exploited in PopPK analysis is using genetic information as an intrinsic covariate. Between-subject variability in a drug’s pharmacokinetics may be due to a myriad of underlying causes, such as age, sex, and genetic differences. Most genetic differences have focused on differences in metabolism, so-called polymorphisms. By definition, a genetic polymorphism is a monogenic trait that is caused by the presence of more than one allele at the same locus in the same population having at least one phenotype (usually defective) in the organism wherein the frequency of the least common allele is more than 1% (Meyer 2001). For example, it is well known that CYP 2D6 has a polymorphic distribution in many different populations, e.g., Caucasians, Asians, etc.

Polymorphisms are usually assessed by either genotyping a subject’s DNA or by phenotyping. Genotyping can easily be determined using samples with genomic DNA (tissue, blood leukocytes, buccal swabs, fingernails, or hair) using polymerase chain reaction (PCR)-based assays. Genotyping has the advantage in that it need only be done once (a genotype is a constitutive property of the individual and does not change over time) and is not subject to outside influences. Alternatively, phenotyping may be conducted which measures the functional expression of a genotype. Examples include the debrisoquine metabolite ratio as a test for CYP 2D6 polymorphism and erythromycin breath test as a surrogate for CYP 3A4 activity. However, phenotyping typically requires the administration of another medication prior to therapy and the results are subject to environmental influence, such as when a concomitant medication is administered. However, since phenotyping is under environmental control, its use may be of more value in explaining within-subject variability. In practice, the use of phenotyping is less common in clinical trials than genotyping and is rarely seen in the PopPK literature.

To date, genotyping has been used solely to explain the between-subject variability in clearance with the genotype treated as any other covariate. Kvist et al. (2001) first studied the role CYP 2D6 genotype plays in the clearance of nortriptyline in 20 subjects with depression and 20 healthy volunteers. CYP 2D6 genotype can be classified into four groups based on the number of functional genes: poor metabolizers (0 genes), heterozygous extensive metabolizers (1 gene), homozygous extensive metabolizers (2 genes), and ultrametabolizers (3 or more genes). Kvist et al. (2001) modeled nortriptyline clearance based on the well-stirred model

$$CL = \frac{Q\times CL_{int}\times \exp(\eta)}{Q+CL_{int}}$$

(49)

where $Q$ was hepatic blood flow fixed at 60 L/h and $CL_{int}$ was intrinsic clearance modeled as a linear function of the number of CYP 2D6 genes (GENE).
The final model for glyburide clearance (CL) was modeled as the sum of the partial clearances

\[ CL = CL_{*1} + CL_{*2} + CL_{*3}, \]

where CL_{*1}, CL_{*2}, and CL_{*3} were the partial clearances if the subject had the *1, *2, or *3 allele, respectively. For example, if subjects had the *1/*3 alleles their clearance would be modeled as CL_{*1} + CL_{*3}. Subjects with the *1/*1 alleles would have their clearance modeled as CL_{*1} + CL_{*1}. Under this model, subjects with the wild-type *1/*1 alleles had a total clearance of 3.5 L/h, but slow metabolizers having a genotype of *3/*3 had a total clearance of only 1.5 L/h.

Zandvliet et al. (2007) present the results of an analysis in 46 Caucasian and 21 Japanese patients who were genotyped for the *2, *3, *4, and *6 polymorphisms of CYP 2C9 and the *2, *3, *4, *5, and *6 polymorphisms for CYP 2C19. Indisulam is a sulfonamide anticancer agent that is well-tolerated but only has limited single agent activity and is currently being studied for the treatment of solid tumors. Phase 1 studies have shown that the dose-limiting toxicity with indisulam was neutropenia and thrombocytopenia. Pharmacokinetically, indisulam is metabolized by CYP 2C9 and CYP 2C19 to form a hydroxylate metabolite that is immediately conjugated to form an O-glucuronide and O-sulfate metabolite. Population analysis showed that clearance could best be described by two parallel pathways, one saturable and the other nonsaturable (Zandvliet et al. 2006). Wide BSV was observed, 45% for \( V_{max} \).

Indisulam pharmacokinetics were extensively characterized in these patients as part of the Phase 1 program. Using their previously reported population pharmacokinetic model as the starting point, the elimination model was expanded to include the genotype as a means to further explain the BSV in indisulam pharmacokinetics. Letting \( P \) be the population parameter in wild-type patients and assuming that homozygous patients will have twice the impact of heterozygous mutations their model for \( V_{max}, K_m \), and linear clearance was

\[ P = P_{pop} (1 - (\theta \times \text{heterozygous} + 2 \times \theta \times \text{homozygous})). \]

Using this model, a 27% decrease in \( V_{max} \) was found between wild-type and CYP 2C9*2 polymorphs and 38% decrease in linear clearance between wild-type and CYP 2C19*3 polymorphs. Patients with these mutations had higher AUC than wild-type patients. Due to the small number of patients a quantitative analysis of CYP polymorphisms and grade of neutropenia could not be made. However, at higher dose levels a relationship between total dose and number of mutations and grade of neutropenia was observed. The pharmacokinetic model was coupled to a cell transit model previously developed for absolute neutrophil count (ANC). Using Monte Carlo simulation, they showed that the relative risk of having dose-limiting neutropenia was 40% higher in patients having a single polymorphism. Homozygous mutations increased risk to twofold. There did not appear to be a

\[ V_{max} = \begin{cases} \theta_1 & \text{if in Group 1,} \\ \theta_1 \theta_2 & \text{if in Group 2,} \\ \theta_1 \theta_3 & \text{if in Group 3,} \\ \theta_1 \theta_4 & \text{if in Group 4,} \\ \theta_5 & \text{if in Group 1,} \\ \theta_2 \theta_6 & \text{if in Group 2,} \\ \theta_2 \theta_7 & \text{if in Group 3,} \\ \theta_2 \theta_8 & \text{if in Group 4,} 
\end{cases} \]

where \( \theta_1 \) was the maximal velocity for the wild-type\(^2 \) metabolizer. \( \theta_2, \theta_3, \) and \( \theta_4 \) were the \( V_{max} \) multipliers if the subject was in the second, third, or fourth group, respectively. \( \theta_1 \) was the \( K_m \) for the wild-type metabolizer. \( \theta_5, \theta_7, \) and \( \theta_8 \) were the \( K_m \) multipliers if the subject was in the second, third, or fourth group, respectively. In this manner, they showed that subjects in Group 4 had 42% lower \( V_{max} \) than the wild-type, whereas \( K_m \) for Groups 2 and 3 were 22 and 54% higher than Group 1, respectively.

Kirchheiner et al. (2002) used another approach to model the clearance of glyburide, an oral hypoglycemic that is metabolized by CYP 2C9, in 21 healthy volunteers. At the time of the study, the genotype for CYP 2C9 consisted of two inherited functional polymorphisms of three different alleles (denoted *1, *2, and *3) that were known to affect the catalytic efficiency of the CYP 2C9 enzyme. Hence, humans were classified as *1/*1 (wild-type), *1/*2, *1/*3, *2/*1, *2/*2, or *2/*3. They could have used a categorical model like Mamiya et al. (2000) but instead chose to model clearance a little differently. The final

\[ CL_{int} = \theta_1 + \theta_2 \times \text{GENE}. \]

Other functional forms to (50) examining the effect of GENE on intrinsic clearance were studied but none of these models resulted in any improvement in goodness of fit over the linear model. The authors also examined debrisoquine metabolite ratio, a phenotypic marker of CYP 2D6 activity, as a covariate but this, too, resulted in no improvement in goodness of fit. Modeling clearance as a function of number of genes resulted in a significant improvement in goodness of fit compared to a two-compartment model without covariates. Using variance component analysis, the authors concluded that CYP 2D6 genes explained 21% of the total variability in the oral clearance of nortriptiline.

Mamiya et al. (2000) later studied the role of CYP 2C19 genotype in the clearance of phenytoin, a drug almost exclusively metabolized by cytochrome-mediated oxidation, in 134 Japanese adults with epilepsy. Phenytoin clearance was modeled using Michaelis–Menten elimination kinetics with parameters \( V_{max} \) (maximal velocity) and \( K_m \) (Michaelis constant). Subjects were classified into four groups based on their genotype. Each group was then treated as a categorical variable and \( V_{max} \) and \( K_m \) were modeled as

\[ V_{max} = \begin{cases} \theta_1 & \text{if in Group 1,} \\ \theta_1 \theta_2 & \text{if in Group 2,} \\ \theta_1 \theta_3 & \text{if in Group 3,} \\ \theta_1 \theta_4 & \text{if in Group 4,} \\ \theta_5 & \text{if in Group 1,} \\ \theta_2 \theta_6 & \text{if in Group 2,} \\ \theta_2 \theta_7 & \text{if in Group 3,} \\ \theta_2 \theta_8 & \text{if in Group 4,} 
\end{cases} \]

\[ K_m = \begin{cases} \theta_1 & \text{if in Group 1,} \\ \theta_2 \theta_9 & \text{if in Group 2,} \\ \theta_3 \theta_9 & \text{if in Group 3,} \\ \theta_4 \theta_9 & \text{if in Group 4,} 
\end{cases} \]

where \( \theta_1 \) was the maximal velocity for the wild-type\(^2 \) metabolizer. \( \theta_2, \theta_3, \) and \( \theta_4 \) were the \( V_{max} \) multipliers if the subject was in the second, third, or fourth group, respectively. \( \theta_1 \) was the \( K_m \) for the wild-type metabolizer. \( \theta_5, \theta_7, \) and \( \theta_8 \) were the \( K_m \) multipliers if the subject was in the second, third, or fourth group, respectively. In this manner, they showed that subjects in Group 4 had 42% lower \( V_{max} \) than the wild-type, whereas \( K_m \) for Groups 2 and 3 were 22 and 54% higher than Group 1, respectively.

\(^2\)Wild-type genes are always designated as the *1 allele and have normal metabolic function.
difference in risk of neutropenia between Caucasians and Asians after controlling for genotype. They then used Monte Carlo simulation to determine the degree of dose reduction needed to equalize the degree of relative risk in patients with high-risk mutations. They found that each CYP 2C19 mutation required a dose reduction of 100 mg/m^2 and that the reduction for a CYP 2C9*3 mutation was 50 mg/m^2. This study beautifully illustrated the role of modeling and simulation in helping to guide dosing regimens in oncology and further illustrates the increasing importance pharmacogenetics may play in personalized medicine.

In testing for whether a genotype effect exists, often the LRT or Wald test is used. Bertrand et al. (2008) presented the results from the first systemic study of model-based test and model selection strategies in determining whether a genetic covariate is important in a model. In their study, they considered a drug whose clearance is controlled by 1 of 3 different genotypes (CC for wild-type homozygotes, CT, or TT).

They used Monte Carlo simulation to test whether analysis of variance (ANOVA) of the EBEs under the base model without covariates, Wald test of the significance of the parameter estimate associated with the covariate, or likelihood ratio test (LRT) comparing the model with and without the genetic covariate was more powerful at detecting a real covariate effect. Also examined was the Type I error rate and model selection approach. Their results were compared to real data from a Phase II trial with indinavir in HIV-positive patients.

Their results indicated that with 40 patients, the Type I error rate was inflated for ANOVA and Wald test using NONMEM’s (Version 5) first-order approximation (FO) and first-order conditional estimation (FOCE) algorithms. The LRT Type I error rate was inflated for FO (47% for 40 patients and 54% for 200 patients) but was near the nominal value of 5% for FOCE (7.9% for 40 patients and 7.9% for 200 patients). With 200 patients, the Type I error rates were near the nominal value for most tests but was still inflated using FO approximation with Wald’s test (10%) and LRT (54%). Only ANOVA had a Type I error rate near the expected level under all conditions studied. The LRT was uniformly most powerful at detecting a significant covariate effect while Wald’s test was the least powerful. The authors also showed that Akaike Information Criteria (AIC) and sample size corrected AIC (AICc) should not be used for model selection in regards to pharmacogenomic studies.

To date, the application of genotyping to PopPK modeling has focused on using this information to explain clearance. Genotyping has not been used to explain the variability in other structural model parameters, such as volume of distribution, but this may change. That polymorphisms exist for the two major drug–protein binding sites in plasma, albumin (Takahashi et al. 1987) and \( \alpha_1 \)-acid glycoprotein (Eap et al. 1988), has been known for some time. Only lately has it been shown that these protein polymorphisms may result in differential protein binding (Li et al. 2002). Hence, in theory, since volume of distribution is dependent on plasma protein binding, those individuals with different polymorphs may lead to polymorphic volume of distribution, a phenomenon that has not been demonstrated in animals or man (yet). And if this is the case, then genotyping may be useful in explaining the variability of other model parameters as well.

While genotyping may seem useful, the models that are developed may only be applicable until a new nomenclature is used or another defective allele is identified. For example, in 2001, when Kirchheiner et al. reported their results with glyburide, only two CYP 2C9 alleles, besides the wild-type, were known. Just 3 years later, there were 12 other alleles besides the wild-type, *2 to *13. How the model developed by Kirchheiner et al. applies to the situation today is difficult to determine and may not be applicable at all. For example, what is the clearance for a subject with the *1/*8 or *1/*5 alleles? Under the model by Kirchheiner et al. there is no answer because these wild types were not included in the model.

It is expected that more and more models will include genotype data as a covariate in the model because evidence indicates that drug development will eventually include DNA microarrays as a routine part of Phase 2 and Phase 3 studies in an attempt to identify responders and nonresponders (Roses 2000). Further, studies are now being done correlating gene expression profiles on gene chips (Affimetrix U95A GeneChip, Santa Clara, CA) to transporter expression in tissue biopsies (Landowski et al. 2003). If during Phase 2 patterns of single nucleotide polymorphisms (SNPs) from DNA microarrays can be found in responders, but not in nonresponders, or gene expression profiles can identify subjects that have abnormal levels of some transporter needed for drug absorption, then the presence of these SNPs can later be used as an inclusion criteria in Phase 3. Hopefully, the probability of “success” for the Phase 3 trial will then be increased compared to had this information not been on-hand. Since, the classification of subjects into groups using gene expression profiles is based on correlative methods, i.e., finding a pattern of SNPs present in responders but not in nonresponders, there is no reason why similar data cannot be used as covariates in a PopPK analysis. To do so, however, result in two problems. One, the number of covariates that would need to be screened would greatly increase since now there may be many hundreds of genomic variables to analyze. Second, many spurious covariates would be added to the model because of multiplicity in hypothesis testing. The benefit of including such genomic covariates in a model, however, is a reduction in the between-subject variability in drug concentrations.
There are times, however, when later it is discovered that a drug’s pharmacokinetics is affected by a genetic polymorphism and that the information related to genotype was not collected as part of a clinical trial. When polymorphisms are present, the distribution of a parameter tends to be wide with heavy tails or may consist of two defined subpopulations. In this instance, a mixture model is frequently used with the hope that the software can classify patients into the correct group, e.g., poor metabolizer or extensive metabolizer. Kaila et al. (2006) used Monte Carlo simulation to see whether NONMEM could correctly classify subjects when the drug’s clearance was polymorphic. The correct classification rate increased when the BSV was decreased, the degree of separation between groups increased, larger sample sizes were used (although this was an inconsistent finding under some scenarios), and the mixing proportion became more balanced, i.e., mixing fractions near 0.5. As a real example they presented the results from two clinical trials in 35 patients taking metoprolol who were phenotyped for CYP 2D6 metabolizer status. They analyzed the dataset in two ways: (1) as if metabolizer status were unknown and CL was modeled as a mixture distribution and (2) using metabolizer status as a covariate on CL in the model. They showed that using the mixture model approach, NONMEM was able to 100% correctly classify patients into the low and high CL group and that the parameter estimates were very similar between the two models. Thus, using a mixture distribution represents a viable approach to model genotyping effects in a population analysis when the genotype or phenotype information was unavailable.

**Incorporating Lag-Times into the Model**

It is not uncommon after extravascular administration for there to be a delay in observing quantifiable drug concentrations. For instance, for an oral drug to be absorbed it must first dissolve into the fluids in the gastrointestinal tract (GI) and then be absorbed in the splanchnic blood whereupon it will then be distributed by the systemic circulation to the rest of the body. The dissolution step may be slow relative to absorption and in fact may be very slow due to either formulation factors or to food stuffs in the GI tract. Such delays in absorption are modeled using a lag-time between the absorption compartment and the observation compartment. There is a subtle difference in how different programs handle lag-times which is important to understand. Programs like WinNonlin and NONMEM model their pharmacokinetics as explicit functions, unless otherwise directed by the user. Lag-times are treated as all-or-none or as a step function. For example, for a drug exhibiting one-compartment kinetics a lag-time would appear in the function as

$$C(t) = \begin{cases} 0 & \text{if } t < \text{lag-time}, \\ \frac{\text{FD}}{V} \left( \frac{k_a}{k_a - k_{cl}} \right) \times \left[ \exp(-k_{cl} \Delta t) - \exp(-k_a \Delta t) \right] & \text{if } t \geq \text{lag-time}, \end{cases}$$

where $\Delta t = t - \text{lag-time}$. Hence, if time is less than the lag-time, drug concentrations will be zero. But once time exceeds the lag-time, concentrations will follow the typical profile of a one-compartment model with absorption but with a shift in the profile to the right. While numerically convenient, the explicit function model is not physiologically realistic.

Another approach to modeling lag-times is to model the kinetic system using differential equations with the lag-time manifested through a series of intermediate or transit compartments between the absorption compartment and observation compartment (Fig. 3). For example, the differential equations for Model A in the figure would be written as

$$\frac{dX_1}{dt} = -k_a X_1,$$
$$\frac{dX_2}{dt} = k_a X_1 - k_{tr} X_2,$$
$$\frac{dX_3}{dt} = k_{tr} X_2 - k_{tr} X_3,$$
$$\frac{dX_4}{dt} = k_{tr} X_3 - k_{t0} X_4,$$

where $k_{tr}$ is called the transit rate constant. Such a series of differential equations does not have an all-or-none outcome and is more physiologically plausible. Using a differential equation approach to model lag-compartments the rise in concentration to the maximal concentration is more gradual. But, as the number of intermediate lag-compartments increases so does the sharpness in the rate of rise such that an infinite number of transit compartments would appear as an all-or-none function similar to the explicit function approach (Fig. 4). Also, as the number of intermediate compartments increases the peakedness around the maximal concentration increases.

Determining the number of transit compartments can be done through trial and error comparing the AIC with each model and then picking the one with the smallest AIC. Alternatively, the LRT can be used as well. However, the computation time increases as the number of differential equations increases leading to run times that may be prohibitive. Savic et al. (2007) presented a slightly different transit model from (56), but both models lead to exactly the
same concentration–time profile (Model B in Fig. 3) after single and multiple dose. In the reformulated version, however, the amount of drug in the \( n \)th transit compartment can be expressed as

\[
X_n(t) = D \left( \frac{k_{tr} t}{n!} \right)^n \exp(-k_{tr} t),
\]

where \( n! \) is the factorial of \( n \). The differential equation for the absorption compartment (\( X_3 \) in this case) can then be written as

\[
\frac{dX_3}{dt} = \left[ D k_{tr} \left( \frac{k_{tr} t}{n!} \right)^n \exp(-k_{tr} t) \right] - k_a X_3.
\]

Hence, in Model B in Fig. 3 the system of differential equations can be reduced to

\[
\frac{dX_1}{dt} = \left[ D k_{tr} \left( \frac{k_{tr} t}{n!} \right)^n \exp(-k_{tr} t) \right] - k_a X_1,
\]

\[
\frac{dX_2}{dt} = k_a X_1 - k_{10} X_2,
\]

\[
C(t) = \frac{X_2}{V}.
\]

The units of \( k_a \) are per time which is not easily interpretable. Instead, mean transit time (MTT), which has units of time, and is equal to

\[
MTT = (n+1)/k_{tr}
\]

can be used instead.

The only difficulty in this equation is finding \( n! \). NONMEM in particular does not support the factorial function, so instead Stirling’s approximation can be used

\[
n! \approx \sqrt{2\pi} \left( \frac{n}{e} \right)^{n+0.5} \exp(-n),
\]

which is accurate to at least \( n=30 \). Savic et al. (2004) used this approach to compare the transit model with the explicit function model using glibenclamide, moxonidine, furosemide, and amiloride and found that the transit model resulted in significantly better goodness of fit than the explicit function model in all four cases.

**Experimental Design Issues**

It is generally recognized that intensive (also called dense or rich) sampling for pharmacokinetic analysis will occur during Phase 1 and, usually, in Phase 2 as well. Pharmacokinetic models obtained from such data are generally considered to be of better quality than models obtained from Phase 3 since intensive pharmacokinetic sampling is not done routinely within a subject in Phase 3. More often, sparse sampling, as few as 1 or 2 samples per subject per dosing interval, is the norm. As one might expect, the ability to obtain accurate and precise parameter estimates in a model derived from sparse data is dependent on the experimental design used to obtain the data. For example, the best times to sample are immediately after dosing and as late as possible thereafter with a one-compartment model after intravenous administration (Endrenyi 1981). For a one-compartment model with absorption it would not be surprising that an accurate estimate of the absorption rate constant cannot be obtained if samples are not collected in the absorption phase. Hence, when samples are collected are of paramount importance. But, balanced against this are usually some restraints on when samples can be collected and how many samples can be collected (usually as few as possible). Sometimes with Phase 3 studies the pharmacokineticist has no say in when samples can be collected and must work with what they are given. It is not uncommon to have Phase 3 studies planned entirely without consultation from a pharmacokineticist and later, at the end of the trial, have the pharmacokineticist analyze the data as part of a regulatory submission.

In cases where the pharmacokineticist has input on when samples can be collected, samples should be obtained at times that maximize the pharmacokinetic “information” about the model parameters while collecting as few as samples as possible. This section will focus on the experimental design considerations to maximize the “information content” obtained during Phase 3 and to survey current practices regarding sample collection times. In order to understand the current state of sample collection and timing for Phase 3 studies it is necessary to first understand the relevant simulation studies on the topic because these are often used to validate clinical decisions on sample times. It must be stated, however, that many of the conclusions drawn from these studies are study specific and do not necessarily generalize to other conditions. For example, the results from a one-compartment model may not apply to two-compartment model.
Fig. 3 Schematic of two different formulations of a one-compartment model with lag-time using two intermediate lag-compartments to model the lag-time. Model A is the typical model, while Model B is the model used by Savic et al. (2004). Both models lead to exactly the same concentration–time profile.

Fig. 4 Concentration–time profiles illustrating the difference between modeling lag-time using an explicit function (55) vs a differential equation approach (56) with a variable number of intermediate transit compartments. Concentrations were simulated using a one-compartment model having a dose of 100 mg, $V = 125$ L, $k_a = 0.7$/h, $k_{10} = 0.15$/h and a lag-time of 3 h. The explicit function models lag-times as all-or-none, whereas the differential equation approach models lag-times more gradually.
Theory Based on Monte Carlo Simulation

Sheiner and Beal (1983) presented the first study on the role of experimental design in one of their seminal papers on nonlinear mixed effects models. They showed that increasing the number of subjects improves parameter estimation accuracy, but that increasing the number of samples per subject does not improve estimation to the same degree when the data were simulated from a one-compartment model. Hence, it is better to have sparse data from more subjects than intensive pharmacokinetic data with fewer subjects. They also showed that relatively accurate and precise parameter estimates (except for residual variance) can be obtained using FO-approximation with as few as 50 subjects having a single sample collected per subject. Keep in mind, however, this was a very simple pharmacokinetic model with only two estimable parameters.

Al-Banna et al. (1990) used Monte Carlo simulation to compare fixed time two and three point designs under a one-compartment model in 50 subjects for a drug having a half-life of 6.93 h. Recall that at the individual level the best times to estimate the model parameters with a one-compartment model after intravenous bolus administration are as early and late as possible. In the two point design, one sample was anchored at 0.08 h with the second sample time varying in 2 h intervals up to 20 h. In the three point design, one sample was fixed at 0.08 h with another at 20 h. The third sample was varied at 2 h intervals up to 18 h. Although not stated, it is assumed the authors analyzed each data set using FO-approximation.

Acceptable bias and precision in the structural model parameters were observed with two samples per subject across any two time points. However, better precision and accuracy in estimating clearance was obtained when the second sample was collected at later times. Volume of distribution was not as affected by the choice of sample times. Across all time points, the variance components were often significantly underestimated with the range of estimates being quite large. When the number of samples per individual was increased to three, regardless of where the middle point was collected in time, the structural model parameters remained unbiased but the bias in the variance components was removed. When the number of subjects was increased to 100 and then 150, the bias and precision in the structural model parameters remained unchanged, but improved the estimation of the variance components. Hence, under these conditions, neither more data per subject nor more subjects improved the estimates of the fixed effects in the model. What were affected were the variance components. Both more data within a subject and more subjects resulted in better variance component estimation.

Breant et al. (1996) followed up the work of Al-Banna, Kelman, and Whiting and used Monte Carlo simulation to determine the number of subjects and samples per subject needed to obtain accurate and precise parameter estimates with a drug that showed monoeponential disposition kinetics. They found that for a one-compartment model, concentration data from 15 to 20 subjects with two samples per subject produced reasonable parameter estimates. Although the authors did not use NONMEM, their results should be applicable to NONMEM analyses.

Booth and Gobburu (2003) examined whether these conclusions held when a single sample was collected at steady-state compared to when multiple samples were collected after a single dose. With only one sample per subject, the bias and precision in clearance was –16.7 and 22.8% for FO-approximation, but was –28.7 and 32.3% for FOCE-I. Paradoxically, the less accurate estimation algorithm decreased bias and increased precision in estimating clearance. The increased bias and decreased precision using FOCE-I to estimate volume of distribution was in general greater than for clearance. But this was not that unexpected since steady-state data contain more information on clearance than volume of distribution. However, FOCE-I was more accurate and precise than FO-approximation (bias: –6.6% for FO-approximation vs. 0.4% for FOCE-I; precision: 6.8% for FO-approximation vs. 3.7% for FOCE-I) when the mix of data consisted of 75% sparse data and 25% dense data. The authors conclude that single trough samples do not provide unbiased and precise parameter estimates even when the exact structural model is known a priori. Unbiased and precise parameter estimates can only be obtained when combined with dense data or when some of the other model parameters are fixed to prior values.

Jonsson et al. (1996) used Monte Carlo simulation to examine the consequences of collecting two samples per visit instead of one per visit when the drug followed a one-compartment model with first-order absorption. Samples were collected after the absorption phase was complete. In the modeling process, they fixed the absorption rate constant \(k_a\) to a constant with all the remaining parameters treated as estimable parameters. Previous studies have shown that for a one-compartment model when no data are collected during the absorption phase, estimation of \(k_a\) is problematic, but when \(k_a\) is fixed to a reasonable value then accurate and precise estimates of the remaining model parameters is still obtainable (Wade et al. 1993).

Jonsson et al. (1996) concluded, as might be expected, that a design with two samples per visit, regardless of when the samples were collected, resulted in better (less bias and more precise) parameter estimates than a design with one sample per visit. The data herein were simulated using a one-compartment model. On a log-scale, concentrations decrease linearly during the elimination phase. Hence, it would be expected that irrespective of when the two plasma samples were collected after absorption was complete, roughly equivalent parameter estimates would be obtained. For more complex models, the timing of the second sample would be crucial in estimating the parameters associated with the additional compartment(s).

Jonsson et al.’s (1996) second conclusion was a logical extension of this argument: two samples are needed to identify more complex models and are especially useful because now interoccasion variability can be estimated.
which is not possible from a one sample per visit design. They also concluded that designs in which some fraction of subjects have only samples early after dosing collected with the remainder of subjects having samples collected later after dosing is inferior to one where most subjects have both early and late samples collected. Hence, balance is important in study design.

Ette et al. (1998) examined the effect of sample size and between-subject variability using balanced study designs wherein an equal number of samples were collected from each subject. In their study, concentration data from 30 to 1,000 subjects were simulated using a two-compartment model with intravenous administration at steady-state. Sample times were chosen based on D-optimality. Once obtained, the concentration–time profile was divided into three windows and two samples per subject were randomly sampled from each time window, i.e., six samples were collected from each subject. Between-subject variability in all the parameters was systematically varied from 30 to 100% with residual variability fixed at 15%. All models were fit using FO-approximation.

In general, accurate and precise estimates of all structural model parameters were obtained across all sample sizes and between-subject variability, except for when the sample size was small (30 subjects) and between-subject variability was large (100%). When between-subject variability was <100%, accurate and precise structural model parameter estimates were obtained irrespective of sample size. In contrast, estimation of the variance components was influenced by both sample size and between-subject variability. Greater numbers of subjects are needed to obtain accurate and precise structural model variance component estimates as between-subject variability increases. However, residual error was always unbiased and imprecise regardless of sample size when between-subject variability was greater than 75%. Hence, any estimates of residual variance with very large between-subject variability should be interpreted with caution.

**Review of Current Practice**

Returning to current practice, the number of samples collected and their timing depend on many variables, few of which are pharmacokinetically based. First, the primary focus of a Phase 3 study is efficacy and safety. If a drug does not demonstrate efficacy, collecting an adequate pharmacokinetic profile for pharmacokinetic analysis becomes moot (although possibly publishable). Similarly, companies are under tremendous pressure to get new drugs to market as soon as possible. Hence, an unstated goal is to complete the trial forthwith. Factors that hinder recruitment, such as intensive pharmacokinetic sampling, are frowned upon by product team leaders and efforts by the pharmacokineticist to collect many samples per subject are discouraged.

Reviewing the population pharmacokinetic literature, focusing solely on Phase 3 studies, the studies can be broadly grouped into three design types. In the first design, blood samples are collected on each visit with the time of last dose recorded. No attempt is made to control when sampling is done relative to dosing. Hence, blood samples are, in essence, randomly collected. The rationale being that the sheer size in the number of subjects will adequately capture the concentration–time profile. In practice, this approach works surprisingly well. For example, Ingwersen et al. (2000) reported this design in the PopPK analysis of 130 epileptic subjects dosed with tiagabine. Subjects were to have a single blood sample collected at eight different visits during an ~1-year period. Sample collection time (relative to dosing) ranged from 0 to 20 h with the bulk of the data obtained within the first 6 h. Visual analysis of the concentration–time profile pooled across subjects clearly showed the data to follow a one-compartment model with an absorption phase.

A modification to this design is when the samples are not collected at random relative to dosing, but are collected at fixed time points across different visits. One convenient choice is to sample at trough, either one sample per subject total or at multiple visits during the study. The FDA Guidance on PopPK (1999) discourages the single trough design since if only a single sample is collected per subject then between-subject variability and residual variability cannot be isolated. With multiple trough samples, although between-subject variability and residual variability can be separated and there is no stated reason not to use the design, the guidance has a general tone of disdain for this design. Further, Booth and Gobburu (2003) have shown that single trough samples do not provide unbiased and precise parameter estimates even when the exact structural model is known a priori.

Another option, the so-called full pharmacokinetic screen, is to sample at fixed times relative to dosing within a single dosing interval such as with a traditional Phase 1 study. For example, Preston et al. (1998) reported such a design in their PopPK analysis of levofloxacin in 313 subjects with bacterial infections of the respiratory tract, skin, or urinary tract. Therein samples were collected at end of infusion, 2, 6.75, 7.75, and 9.25 h after the third intravenous dose. What made intensive sampling possible in this case was that these subjects were at the hospital already. In general, if the disease state requires hospitalization then intensive sample collection is less a problem than if the subject is treated in an out-patient manner.

The second design is to use a time window wherein blood samples are collected anytime within specified intervals after dosing either within a visit, or more likely, across multiple visits. One recommendation is to collect three samples per subject, one of which is at trough (Aarons et al. 1996). For example, for most orally absorbed drugs the time to maximal concentration is 2–4 h after administration. Assuming that drug concentrations can only be collected during a visit to the physician’s office, a three point design may be predose (0 h), 2–4 h, and 4–8 h after dosing once steady-state is reached. To obtain an estimate of interoccasion
variability this sampling schedule would be repeated the next time the subject returns to the physician’s office.

The pharmacokinetic window design has many variants. One variant is where each subject is sampled twice on each visit. The subject comes to the clinic, has a blood sample drawn at predose, the next dose is taken, the subject remains at the clinic for some time, and then a second sample is collected after a period of time has expired. The disadvantage of this approach is that it requires the subject to remain at the clinic, which places an extra burden both on the subject and on the site in collecting the sample. The advantage of this design is that the quality of the dosing times and sample collection times is usually better than other designs because a third-party (the study nurse or phlebotomist) records the sample and dose times. For example, Jackson et al. (2000) reported such a design in the PopPK analysis of nelfinavir mesylate in human immunodeficiency virus-infected (HIV) patients. Therein, blood samples were to be collected at each visit at trough and at postdose 2 h later, with most samples collected on Weeks 2 and 8 of the study.

Another modification of this design is that on half the visits a blood sample is collected at trough, while the remainder of visits the sample is collected sometime after dosing. Hence, each subject has a single sample collected per subject per visit. For example, Phillips et al. (2001) used such a design in the PopPK analysis of azimilide. Sample times ranged from 0 to 36 h after dosing with two groups of data, one within 0 to 12 h after dosing and another from 23 to 32 h after dosing. One disadvantage of this approach subjects are required to either keep a dosing diary or to remember the time of their last dose, both of which are often not of the best quality. It is not uncommon to review a diary log on a subject and see that all daily doses were taken at the same time of day, say 8:00 PM. Were the doses really taken at that time or was this “about” the time the doses were taken and the subject simply rounded to the nearest hour? With a PopPK analysis, often done after the study is complete, it is difficult to go back and re-assess the quality of this data. Another disadvantage of this approach is that no estimate of interoccasion variability can be made.

The last design type is where some fraction of subjects undergo intensive pharmacokinetic sampling, usually at a few sites, while the remaining subjects have sparse sampling throughout the study. For example, Moore et al. (1999) reported such a design in their PopPK analysis of lamivudine in subjects infected with HIV. Therein select centers were to collect six samples per subject per visit at fixed times over an 8-h period. The remaining subjects had two samples per visit collected. One sample was taken when the subject arrived at the clinic. If the last dose taken was within the last 6 h, the next sample was taken at least 1 h after the first sample. If the last dose was not taken within the last 6 h, the subject was to take a dose of medication, and have the sample collected at least 1 h later.

One issue that inevitably arises is when to collect the samples in the fixed time point design. One method is to use D-optimality or one of its variants to identify the time points to collect samples. These methods are numerically intensive and beyond the ability of the average pharmacokinetic modeler. The reader is referred to Retout et al. (2002), Ette et al. (1994b) and Duffull et al. (2002) for further details. Another method is based on computer simulation. One can simulate data sets having various design considerations and see which one gives the best estimates of the model parameters (Kowalski and Hutmacher 2001). Of course, both D-optimality and simulation based approaches assume that previous knowledge of the pharmacokinetics of the drug are known. Sometimes this may not be the case. While D-optimality designs have been used in practice (Karlsson et al. 1998; Nguyen et al. 2002) it is unclear whether these methods offer any practical advantage over pharmacokinetic screens or pharmacokinetic windows.

On the Detection of Subpopulations

Lee (2001) used Monte Carlo simulation to examine the influence of experimental design factors on detecting a pharmacokinetic difference in a subpopulation of individuals having 30% higher clearance than the main population. The factors examined included number of subjects in the subpopulation, sampling scheme, and degree of compliance. Subjects were simulated having twice-a-day dosing with perfect compliance. Concentration–time data were simulated using a one-compartment model with first-order absorption with two samples per subject collected at steady-state. In most cases, the samples were collected 1 h after dosing and 11.5 h after dosing. Lee did not use NONMEM for his analysis but instead used S-Plus’s (Insightful Corp., Seattle, WA) NLME function, which uses a FOCE algorithm to estimate the model parameters. Clearance was modeled as

$$CL = \theta_0 + \theta_{slope}G,$$  \hspace{1cm} (62)

where $\theta_0$ is the clearance in the main population, $\theta_{slope}$ is the shift in clearance in the subpopulation from the main population, and $G$ is a dichotomous covariate coding whether a patient is in the main group ($G = 0$) or subpopulation ($G = 1$).

In most scenarios, 100 subjects were simulated with the subpopulation having a clearance 30% higher than the main population. One exception was the simulation designed to detect a false subpopulation (Type I error rate) in which $\theta_{slope}$ was set equal to zero. The presence of the subpopulation was tested using two different methods. The first method was the LRT comparing the full model (62) to the reduced model where the term $\theta_{slope}G$ was not included in the covariate model for clearance. The second method was a $T$-test on $\theta_{slope}$

$$T = \frac{\theta_{slope}}{SE(\theta_{slope})},$$ \hspace{1cm} (63)

where $SE(\theta_{slope})$ is the standard error of the estimate for $\theta_{slope}$. For each study design combination, 200 replicates were simulated. The power of each method (LRT and $T$-test) was estimated as the number of times the null
hypothesis of no subpopulation was rejected using a significance level of 0.01 divided by the total number of simulations.

Under the basic design, the Type I error rate for the $T$-test and LRT was 2 and 1%, respectively, using a significance level of 1%. The power to detect the subpopulation increased for all methods when the proportion of subjects in the subpopulation increased. A total of 80% power was reached for the LRT when 20% of the population was in the subgroup and never reached 80% for the $T$-test. Power did not increase much for either method when the percent of subjects in the subpopulation was increased to 30%. The LRT had greater power at detecting a subpopulation than did the $T$-test under all conditions studied.

Many modifications to the basic scenario were also examined. In the first modification, no effect on overall power was observed when the total number of subjects was doubled from 100 to 200 (but keeping the percent of subjects in the subpopulation the same). In the second modification, no difference in power was observed when the sampling times were allowed to vary randomly by ±10% from their fixed values. In the original design, samples were collected at 1 and 11.5 h on Day 10. In the third modification, an additional sample was collected at 5 h postdose. Power to detect a subpopulation increased dramatically with 80% power being achieved with only ten subjects in the subpopulation for both the LRT and $T$-test. Hence, three samples was much better than two samples.

The timing of the samples also appeared critical. When the two steady-state samples were collected at trough, instead of at 1 h postdose, and at 11.5 h postdose, the design resulted in a significant decrease in power. Maximal power reached only ~40% for both the LRT and $T$-test even with 50% of the population consisting of the subpopulation. But when 25 of 100 subjects had a complete pharmacokinetic profile collected at 1, 3, 5, 8, and 11.5 h postdose and the remainder of subjects were sampled twice at trough, 80% power was achieved using a $T$-test when 30% of the population was in the subpopulation. In this instance, the only instance when this was true, power was never more than 40% using the LRT. Interestingly, including the full profiles with trough samples also increased the false positive rate to as much as 25%. This high false positive rate remained near 15% when instead of two troughs one of the samples was collected 1 h after dosing at steady-state.

Lee’s study clearly showed that experimental design factors can have an enormous impact at detecting subpopulations within a population. Increasing the number of subjects in a population from 100 to 200 did not appear to have much impact as long as the size of the main population was much larger than the size of the subpopulation. Much more important was the number of samples collected per subject and the timing of those samples. Good power at detecting a subpopulation and good accuracy at estimating the magnitude of the difference in clearance between the main group and subpopulation could be had with as few as two samples per subject collected at steady-state but was dependent on when those samples were collected. The worst power was obtained if the samples were collected only at trough. Adding complete pharmacokinetic profiles from a proportion of subjects with the remainder having only trough samples did not improve power using the LRT but did so using a $T$-test. Surprisingly, adding subjects with complete profiles to subjects with only trough data collected also increased the false positive rate to unacceptable high values.

Holford (2002) later repeated Lee’s experiment (which used S-Plus’s NLME function) with NONMEM using FOCE-I and found that the power to detect a subpopulation was substantially worse for all the designs compared to the results reported by Lee. Like Lee’s results, trough samples were the worst design overall. Variable sampling designs where the samples were collected within a time window made the results more stable and less biased but, in contrast to Lee, did not improve the power of the design. It is unclear the reason for the discrepancy between studies, but may have been due to differences between the estimation algorithms within the two software packages.

The presence of subpopulations in a parameter, like the polymorphic distribution of clearance, are sometimes assessed using a mixture model. Under this approach, a parameter is defined as consisting of two groups and a mixing proportion, e.g., Group 1 has a clearance of 15 L/h and Group 2 has a clearance of 5 L/h with 80% of the patients being in Group 1. The software then determines whether a patient is best suited for Group 1 or Group 2. Kaila et al. (2006) studied how well this approach works in practice using a combination of Monte Carlo simulation and a real example using metoprolol. The ability to distinguish groups increased when the BSV was decreased, the degree of separation between groups increased, larger sample sizes were used (although this was an inconsistent finding under some scenarios), and the mixing proportion became more balanced, i.e., mixing fractions near 0.5. Using real data from two clinical trials in 35 patients taking metoprolol who were phenotyped for CYP 2D6 metabolizer status, they analyzed the dataset two ways: (1) as if metabolizer status were unknown and CL was modeled as a mixture distribution and (2) using metabolizer status as a covariate on CL in the model. They showed that using the mixture model approach, NONMEM was able to 100% correctly classify patients into the low and high CL group and that the parameter estimates were very similar between the two models. Thus, using a mixture distribution to isolate subpopulations is a viable approach.

**General Guidelines for Sample Collection**

So what are the general guidelines for sample collection in a Phase 3 study? It is better to collect sparse data from more individuals than intensive data from few individuals. Within any dosing interval at least three to four samples per subject should be collected. Whenever possible collect at
least three samples on more than one dosing interval to estimate interoccasion variability. Collect samples within a visit in a balanced manner. It is better to have the majority of subjects have samples collected early after dosing and later after dosing than to have some fraction of subjects with all their data early after dosing with the remaining subjects having all their data late after dosing. Whether to use a pharmacokinetic window approach or to use a pharmacokinetic screen where sample times are random is arbitrary and depends on the complexity of the study and the burden to the site and subject. Lastly, in the author’s interactions with the regulatory agencies, study designs where a fraction of the subjects have intensive pharmacokinetic sampling and the remainder have sparse sampling are favored over purely sparse sample designs.

In summary, many experimental designs have been used in the collection of Phase 3 pharmacokinetic data. None of them have been proven to be superior in practice. In theory, D-optimal designs offer more efficient parameter estimation over pharmacokinetic screens and windows, but this has not been demonstrated in practice. Clearly, D-optimal designs represent a hardship on a Phase 3 study that few companies are willing to undertake, without perhaps some prodding by regulatory authorities. For example, a D-optimal design may call for a sample to be drawn 6.5 h postdose, which may be a burden on a patient to come in at that particular time or to remain at the clinic after dosing for collection of that sample. Until evidence is provided on the superiority of these designs, pharmacokinetic screens and windows will continue to be commonplace.

Toxicokinetic Analyses with Destructive Sampling

As defined by the International Conference on Harmonisation (1995), toxicokinetics (TK) is the “generation of pharmacokinetic data, either as an integral component in the conduct of nonclinical toxicity studies or in specially designed supportive studies, in order to assess systemic exposure.” The usual battery of toxicity studies to support an Investigational New Drug (IND) application include single dose and repeated dose toxicity studies in two species (rodent and nonrodent), at three or more dose levels, in both males and females, by the proposed route of administration. The usual species studied are rat and dog. With dogs, repeated sampling within an animal is not problematic given the typical blood volume in a dog. With rodent species, repeated sampling within an animal is problematic because of the limited blood volume and difficulty in the collection of repeated blood samples. Also, repeated sampling increases the stress of the animal and may confound the toxicologic results. Hence, a typical TK study will employ “satellite” animals, animals that are dosed exactly the same as the other animals in the study, but whose sole purpose is to be used for TK assessment.

Even with satellite animals, repeated sampling within a rodent species is often not performed. Usually, each animal contributes to the data at a single time point. When the animal is sacrificed to obtain the pharmacokinetic data, these animals are said to be destructively sampled. The number of animals in such studies is large and costly since often a study has many different dose levels, are dosed in both males and females, and samples are collected at many time points. As a compromise between cost, number of animals, and manageability, the usual TK study is sparsely designed with concentration data collected at 5–7 time points over a dosing interval with 3–5 animals/sex/dose per time point. As might be expected, due to the sparseness of the design, the choice of sample times in such studies is critical to obtaining valid parameter estimates. Not sampling long enough may miss a phase in the disposition profile of the drug, leading to an underestimate of area under the curve, and an overestimate of clearance. On the other hand, sampling too long may lead to concentrations that are below the lower limit of quantification of the assay and may not be usable. It is beyond the scope of this section to discuss proper sample collection times in TK studies, suffice to say that proper sample times are paramount for accurate parameter estimates. The reader is referred to Smith et al. (1990) and Ette et al. (1995) for further details.

The primary TK endpoints from toxicity studies include measures of exposure, such as area under the curve and maximal concentration, time to maximal concentration, and half-life. Usually, these metrics are obtained by calculating the mean or median concentration at each time point and then using noncompartmental methods to obtain the pharmacokinetic parameter estimates. The problem with this approach is that it ignores between-subject variability. The variability between animals is lumped into a “single” animal and the estimate of the parameter is obtained. In doing so, the error associated with the parameters is biased because of the ignored between-subject variability.

Many simulation studies have shown that nonlinear mixed effects modeling of destructively obtained TK data results in relatively accurate and unbiased estimates of the theoretical pharmacokinetic parameters (Ette et al. 1994a; Jones et al. 1996) that are more accurate than naive pooling and noncompartmental analysis (Hing et al. 2001b). The phrase “relatively accurate” is used since the sampling times and data are so sparse that very accurate estimation of population parameters is rarely achieved. In a study where more than one sample is collected per subject, the total variability in the concentration data can be isolated into its components: between-subject, intrasubject, or interoccasion (if the animal is sampled on more than one dosing occasion), and residual variability. With only one sample per subject, the components of variance cannot be identified. If residual variability is assumed negligible then the only variance component that can be obtained is between-subject variability. This assumption is reasonable since between-subject variability is usually an order of magnitude greater than residual variability (Lindstrom and Birkes 1984). Another approach is to fix residual variability to some constant, usually the variability of the assay, and allow between-subject variability to remain an estimable
parameter. Hing et al. (2001b) showed that setting residual error to near zero, equal to assay error, or twice assay error had little effect on the values of the estimated model parameters.

An early example of this methodology was presented by Burtin et al. (1996) who presented the results of a PopPK analysis from a 13 week toxicology study in male and female rats orally dosed once daily at four dose levels. Each animal provided one sample on the first day of dose administration and one sample after the last dose on Day 92 at one of five possible times (0.5, 1, 2, 7, and 24 h postdose) at the same time on each occasion. Two animals were sampled at each of the five times. They then compared the results of the PopPK analysis to a traditional noncompartmental approach. Both analysis methods came to similar conclusions, but the PopPK approach, resulted in greater mechanistic interpretations to the data.

TK studies fall under the more broad category of sparse data analysis, which may arise from many different settings, such as when tissues are sampled in addition to blood. For example, Aarons (1991) reported using PopPK methods to characterize the elimination kinetics of flurbiprofen in synovial fluid in 26 rheumatoid arthritis who had only a single synovial fluid sample collected. While it is clear that the use of nonlinear mixed effects modeling of destructively sampled TK data can lead to reliable population estimates and good individual parameter estimates that can be used to generate measures of exposure, the use of this approach in routine drug development has not been implemented, possibly due to the small number of toxicologists who are familiar with the methodology and whether the time invested in the increased complexity of the analysis leads to true benefits in clinical drug development.

**Handling Missing and Censored Data**

The reader is advised that before beginning this section to read the section on missing data in the chapter on Linear Models and Regression. The issue of missing data is so complex that entire books have been written on the subject, most notably Little and Rubin (2002) and Schaefer (1997). As such it is beyond the scope of this book to cover all aspects of handling missing data. Nevertheless, a brief review of the issues involved and current strategies for dealing with missing data in the longitudinal setting will be discussed. The case where the observations are independent, such as in a linear regression model, has already been covered. What raises the issue again in this chapter is the possibility that observations are correlated within subjects. In other words, all the observations in a linear regression problem are independent and any of the techniques described in the chapter on Linear Models and Regression may be used. But with a mixed effects model the data are hierarchical and nested within individuals and as such those techniques that are based on the independence of the data are no longer valid.

In a random review of 20 PopPK papers published between 2000 and 2002 in leading journals, e.g., British Journal of Clinical Pharmacology, not a single paper even mentioned whether any data were missing, yet alone how missing data were handled. Nevertheless, despite the absence of these techniques in the PopPK literature it is still important to discuss their use and utility so that more widespread use will be seen in the future. It is likely that these techniques are not seen in the PopPK literature simply because most analysts are unfamiliar with them.

**When the Dependent Variable is Missing**

When the dependent variable is missing, the first question to ask is “why are the data missing”? If the dependent variable is missing completely at random (MCAR) then the missing data is ignorable and the usual solution is to use a complete case analysis. Neither linear or nonlinear mixed effects models are penalized when the missing dependent variable is simply deleted from the analysis – the resulting model parameters are unbiased but may not be as precise due to a decrease in the sample size.

The real issue with a missing dependent variable is if the dependent variable is missing because of the value of the measurement, in which case the missing data are “nonignorable.” For example, the concentration of drug in a sample may be below the lower limit of quantification (LLOQ) of the assay in which case the exact value of the measurement is unknown, but it is known that its value is less than the LLOQ. Such data are said to be “left-censored.” One method to handling left-censored data is deleting the censored observation. If the data follow a one-compartment model, deleting left-censored data from the analysis has little impact on the model parameter estimates (Hing et al. 2001a), which makes sense because the concentration–time profile is declining in a log-linear manner. Linear extrapolation of the observed data to values below the LLOQ on a log-scale would not be expected to be that different from observed data had a lower LLOQ been available. However, omission of left-censored data is not recommended for multiphasic concentration–time profiles. Duval and Karlsson (2002) showed that severe parameter bias can occur when left-censored concentration data are omitted from a pharmacokinetic analysis in the case of a two-compartment model. Using simulation, Duval and Karlsson showed that clearance tended to be underestimated, peripheral volume was overestimated, and that terminal half-life tended to be overestimated as a consequence. Distributional parameters were not affected to any significant degree since most of the information used in estimating these parameters occurs in the early portion of the concentration–time curve. The degree of bias in the parameter estimates was related to many factors, including the proportion of missingness, the shape of the concentration–time profile, the dosing regimen, and sampling scheme.
Fig. 5 Scatter plot of concentration–time data simulated under a one-compartment model (top). The limit of quantification of the assay (LLOQ) was 100 ng/mL (gray line). The solid line and dashed line shows the best nonlinear least-squares fit using a one- and two-compartment model, respectively, both ignoring the censored data. Ignoring the censored data, the presence of observations above the LLOQ tends to drive the model to the more complex one because concentrations above the LOQ late in the time profile give the appearance of multiphasic kinetics.

Byon et al. (2008) examined how censoring impacted the choice of a pharmacokinetic model, in this case a one- or two-compartment model. Their results showed that based on the LRT, with a small amount of censoring, the two-compartment model is more frequently chosen when the data are treated as missing, even when the true model is a one-compartment model, because these observations are ignored during the fitting process (Fig. 5). These results suggest that simply ignoring censored observations, even when the percent missing is low, can lead to a wrong choice in the pharmacokinetic model.

Most often, left-censored data are imputed using some fraction of the LLOQ because of the ease in its implementation. One suggestion to handle missing left-censored data is to replace the first missing value with one-half the LLOQ and then delete all missing values thereafter. Of course, this suggestion assumes that the time distance between the last observed concentration and missing observations is small relative to the drug’s half-life. To account for the imputed value the residual error is corrected by the addition of a constant term equal to one-quarter the value of the LLOQ, which assumes that the data around one-half the LLOQ is normally distributed. However, Duval and Karlsson tried this method in their analysis and showed that this method did not improve the model parameter estimates. A modification of this method is to set the missing values to zero, but such a substitution may lead to infinite weights and problems in the optimization process. Hence, zero substitution cannot be recommended. The use of one-half the LLOQ is not tied to any specific theory. Indeed, any value between 0 and the LLOQ could be used, even possibly random draws from a uniform distribution on the interval [0, LLOQ).

To truly account for left-censored data requires a likelihood approach that defines the total likelihood as the product of the likelihoods for the observed data and the missing data and then maximizes the total censored and uncensored likelihood with respect to the model parameters. In the simplest case with \( n \) independent observations that are not longitudinal in nature, \( m \) of which are below the LLOQ, the likelihood equals

\[
L = \prod_{i=1}^{m} p(Y_i < \text{LLOQ}) \prod_{i=m+1}^{n} p(Y_i = Y_i).
\]

It should be noted that in the case of right-censored data the likelihood is simply

\[
L = \prod_{i=1}^{m} p(Y_i > \text{ULOQ}) \prod_{i=m+1}^{n} p(Y_i = Y_i).
\]

where ULOQ is the upper limit of quantification. Lynn (2001) illustrated this approach in modeling the distribution of HIV RNA concentrations with left-censored observations and compared the results with substitution methods using the LLOQ and one-half the LLOQ. Lynn showed that when 15% of the observations were missing, simple substitution methods performed well relative to maximum likelihood (ML) or multiple imputation (MI) methods, but that substitution methods did not perform as well with heavier censoring.

With mixed effects data where the data are longitudinal in nature the likelihood becomes more complex and difficult to implement. Beal (2001) present some solutions to left-censored missing data within the context of a nonlinear mixed effects model. He proposed the following solutions:

- Method M1: discard BQL observations and analyze only those samples above the LLOQ using extended least squares
- Method M2: Discard BQL observations and use conditional likelihood to estimate the model parameters
- Method M3: Maximize the likelihood taking into account left-censored observations
- Method M4: Like Method M3 but the likelihoods for observations above and below the LLOQ are conditioned on the observations being greater than zero
- Method M5: Replace all BQL observations by LLOQ/2 and apply extended least squares
• Method M6: Replace the first BQL value with LLOQ/2 and discard the rest of them as in Method M1
• Method M7: Replace the first BQL observation with 0 and discard the rest of them

Using Monte Carlo simulation, Beal examined these methods and concluded, not surprisingly, that the substitution methods did not perform as well as the ML methods. Method M7 resulted in the greatest bias. Under certain conditions, no difference between the substitution methods and ML methods was observed but were situation specific. Under the right conditions, the substitution methods can be shown to fail spectacularly compared to the ML methods. As such, the ML methods are recommended when left-censored data are present. The problem with the ML methods is that they are not easily implemented in any software package.

Ahn et al. (2008) extended the simulations of Beal and concluded that for overall performance Method M3 was the best followed by M2 and then M1. Method M3 also resulted in fewer problems with nonsuccessful terminations than the other methods. Method M3 directly uses (64) as the likelihood and requires use of the Laplacian option within NONMEM for parameter estimation. As will shown in later chapters, Method M3 is the approach used to handle censored data in survival analysis. Assuming the residuals in later chapters, Method M3 is the approach used to handle censored data in survival analysis. Assuming the residuals are normally distributed with mean 0 and variance $g(t)$, the likelihood for an uncensored observation $Y(t)$ is

$$L_{\text{obs}}(t) = \frac{1}{\sqrt{2\pi g(t)}} \exp \left[ -\frac{(Y(t) - f(t))^2}{2\sigma^2} \right], \quad (66)$$

where $f(t)$ is the model predicted value. The likelihood for a censored observation is based on the cumulative normal distribution $\Phi$

$$L_c(t) = \Phi \left( \frac{LLOQ - Y(t)}{\sqrt{g(t)}} \right). \quad (67)$$

The total likelihood is then

$$L = \prod_{i=1}^{n} L_{\text{obs}}(t) \prod_{i=m+1}^{n} L_c(t). \quad (68)$$

Beal also presented a conditional likelihood (Method M2) where the likelihood is conditioned on the observation being above the LLOQ. In other words, the observations are treated as arising from a truncated distribution where the lower limit of truncation is the LLOQ. Therefore, for only concentrations above the LLOQ the likelihood is

$$L = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi g(t)}} \exp \left[ -\frac{(Y(t) - f(t))^2}{2\sigma^2} \right] \left[ 1 - \Phi \left( \frac{LLOQ - Y(t)}{\sqrt{g(t)}} \right) \right]. \quad (69)$$

These concepts are illustrated in Fig. 6. Implementation of Methods M2 and M3 is easier in NONMEM Version 6 and higher than previous versions using the YLO option (Method M2) and F_FLAG option (Method M3). F_FLAG is an option in NONMEM that allows a model to be specified by a predicted value for some observations and a likelihood for others, whereas YLO is a lower bound for the truncating distribution. F_FLAG is also useful for simultaneous modeling of categorical and continuous data, where the categorical data are defined by a likelihood and the continuous data are defined by a predicted value, like a concentration. Control streams illustrating these methods are presented in Ahn et al. but will need to be modified if the residual distribution is not assumed to be normal.

**When the Independent Variable is Missing**

Missing covariate values cannot be ignored in an analysis because many software packages do not allow missing independent values. With some software packages, subjects with missing covariates are either completely ignored, i.e., that subject is not used in the analysis, or the missing values are set equal to zero. One way to handle missing covariate information is to delete the observation from the analysis, the so-called complete case approach. This approach tends to be most useful when the sample size is large and a small fraction of the covariate is missing. Complete case analysis does not result in biased parameter estimates but simply acts to decrease the precision of the estimates by decreasing the sample size. More often, however, substitution methods are used whereby an imputed value is substituted for the missing value and the analysis proceeds as is the data were never missing in the first place. If the missing covariates are static over time, such as if all the covariates are fixed baseline laboratory values, then any of the imputation methods presented in the chapter on Linear Models and Regression can be used. If,

![Fig. 6: Illustration of the concepts of Methods M2 and M3. Method M3 maximizes the likelihood of both the observed and censored observations, whereas Method M2 maximizes the likelihood of observations above the LLOQ conditional on them being above the LLOQ, i.e., they are from a truncated distribution. Figure redrawn from Bergstrand et al. (2007) with permission](image-url)
however, the covariates change over time within a subject then these methods are no longer valid as they fail to account for correlations within an individual.

In dealing with missing covariate data that are time-dependent, it will be assumed that excluding all subjects with missing data is not an option, that the subject has at least one covariate observation available, and that somehow the missing data must be imputed. Suppose total bilirubin concentration in serum, a marker for hepatic function, was assayed at Weeks 1 and 4, whereas pharmacokinetic samples were collected at Weeks 1–4. Hence, there is no total bilirubin concentration available at Weeks 2 and 3. The data are not simply missing. The data were never collected. Since there is some information on the missing values, albeit at different time points, then one option is to impute the missing value based on the observed values. A common imputation method is carry-forward analysis, sometimes called last-observation carried forward (LOCF), where the missing value is set equal to the value of the last observed value. So, in this case the total bilirubin concentration at Weeks 2 and 3 will both be set equal to the concentration observed on Week 1. The data are then analyzed using all the data (observed plus imputed). It is assumed under carry-forward analysis that the distribution of values at the time the data were missing is the same as the last time the value was observed and that the covariate remains constant over time. As the time distance between measurements increases, these assumptions becomes more and more untenable.

Along with carry-forward imputation are a slew of modifications including interpolation, subject-specific mean value imputation, subject-specific regression imputation, and normal value imputation. Suppose that total bilirubin was 0.4 mg/dL on Week 1 but 1.0 mg/dL on Week 4. Should the values on the second and third weeks be the last observation carried forward (LOCF), i.e., 0.4 mg/dL, or should some type of interpolation between 0.4 and 1.0 mg/dL, such as 0.6 mg/dL on Week 3 and 0.8 mg/dL on Week 4, be used? Little to no research has been done examining this issue in the context of PopPK analyses. In this instance, it makes sense that some type of interpolation will be of better predictive value than LOCF, especially when the values drastically change between visits. Albridge et al. (1988) showed that interpolation of missing data when the values are time-dependent generally results in smaller relative errors from true values than other methods, such as LOCF, naivey inserting the subject mean, or linear regression of correlated co-recorded covariates (which is what the FDA Guidance on PopPK suggests). In this example, bilirubin values at Week 2 could be linearly interpolated using the two bracketing observed concentrations with the equation

\[ \hat{x}_2 = x_1 + \frac{x_4 - x_1}{t_4 - t_1} (t_2 - t_1), \]  

where \( \hat{x}_2 \) is the interpolated value at Week 2, \( x_1 \) is the observed bilirubin concentration at Week 1, \( x_4 \) and \( x_1 \) are the observed bilirubin concentrations at Week 1 and 4, respectively, and \( t_i \) is Week \( i \). So the interpolating equation for Week 2 would be

\[ \hat{x}_2 = 0.4 + \frac{1.0 - 0.4}{4 - 1} (2 - 1) = 0.6 \text{ mg/dL}. \]  

Higher-order interpolating polynomials can be developed and should be considered if the covariate changes curvilinearly over time. The reader is referred to Chapra and Canale (1998) for further details on other interpolation methods, such as Lagrange interpolating polynomials. One problem with interpolation is that the interpolation may not accurately reflect the covariate pattern over time and may not be as efficient if only data near the missing value are used (Higgins et al. 1997).

With subject-specific mean value imputation, the data within an individual are averaged over time (sometimes with weights proportional to the time-interval between measurements) and then the average value is substituted for the missing value. Again, the average value may not be reflective of the true value if the covariate is changing over time. Subject-specific regression imputation uses a model, either parametric or semiparametric, to predict the value of the missing observation and may be more useful than the other within-subject methods as it uses all the data within an individual and can account for time trends. The predicted value may be model-dependent, however, and a poor model may produce biased imputed estimates. Substituting normal values or disease-specific “normal” values cannot be advocated because the method completely ignores the population of interest and the individual in particular. All the imputation techniques just discussed are useful because they are easy to implement and can be easily communicated to nonmodelers. But, these techniques are all flawed in that they do not take into account the uncertainty in the estimate of the imputed value, a criticism which may be alleviated by using the multiple imputation techniques of Rubin (1987).

The within-subject imputation methods just described tend to be more useful when there is a lot of data within an individual. When the data are sparse, the imputation methods are suspect. A more useful method would be one where the imputation is not only based on within-subject observations but between-subject observations. Thus, although information is used at the individual level, information is also used across individuals. Wu and Wu (2002) developed such an imputation scheme using the Gibbs sampler to impute missing data in the kinetic modeling of RNA viral load in patients with human immunodeficiency virus (HIV). Using Monte Carlo simulation, Wu and Wu compared their Gibbs sampler approach to within-subject mean value (MV) imputation, LOCF imputation, and complete case (CC) analysis. The Gibbs sampling approach produced the smallest bias and had the greatest precision of the methods studied. The other methods had significantly larger bias and less precision than the Gibbs sampling approach. When these methods were applied to real data, the LOCF and MV approaches produced some parameter estimates with completely different signs than the MI approach. For example, one
Nonlinear Mixed Effects Models: Practical Issues

model parameter produced by CC and MV was positive, whereas the MI and LOCF estimate were negative. Because of the large bias produced by the other methods, Wu and Wu proposed that the parameter estimates produced by MI were the most reliable. However, the Gibbs sampling algorithm is not implemented in any software currently and was designed by Wu and Wu for their problem. Hence, routine use of this method cannot be done at this time.

Sometimes, if the missing covariate is censored and not ignorable in the reason for missingness, rather than treating the missing covariate as a continuous variable, the covariate can be categorized and the new categorical variable used in the analysis instead. For example, in a PopPK analysis of gentamicin in 210 subjects with cancer, Rosario et al. (1998) created three new covariates based on serum creatinine concentrations and different “threshold” criteria. With Covariate 1, all creatinine concentrations less than or equal to 60 μmol/L (the lower limit of the reference range used by the hospital) were set equal to 60 μmol/L. All other samples were set equal to their original value. With Covariate 2, all creatinine concentrations less than or equal to 70 μmol/L were set equal to 70 μmol/L (an arbitrary intermediate value). All other samples were set equal to their original value. In Covariate 3, all creatinine concentrations less than or equal to 84 μmol/L were set equal to 88.4 μmol/L, the equivalent of 1.0 mg/dL. All other samples were set equal to their original value. All three categorical models resulted in a significant improvement in goodness of fit compared to the model treating serum creatinine concentration as a continuous variable.

If the data are collected at fixed time-intervals then one trick to generate imputed values that would account for within-subject correlations is to transform the data into a columnar format with one row of data per subject. So, if the data were collected at Visits 2–4, then three new variables would be generated. Variable 1 would correspond to Visit 2, Variable 2 to Visit 3, etc. In this manner then each row of data would correspond to a single individual. Now any of the imputation techniques introduced in the chapter on Linear Regression and Modeling could be used to impute the missing data based on the new variables. Once the data are imputed, the data set can be reformatted to multiple rows per subject and the analysis proceeds with the imputed data. This approach assumes that all samples are collected at the same time interval for all subjects, i.e., assumes that all samples were assumed at Visits 1–4 in this case.

Because of the need for easy to implement approaches to handling missing data, pharmacokinetic modelers tend to use either simple imputation or a dummy variable approach. An example of the dummy variable approach is given by de Maat et al. (2002). A binary missing value indicator, denoted as MISS, was created which was set equal to “0” if the value was not missing and “1” if the value was missing. Then to model the typical clearance value (TVCL), for example, with a dichotomous covariate $x$ (taking values 0 and 1) having some proportion being missing, the covariate model was defined as

$$TVCL = \theta_1 + \theta_2 x^{(1 - MISS)} \theta_3^{MISS}. \tag{72}$$

Under this model are four possible outcomes

<table>
<thead>
<tr>
<th>$x$</th>
<th>MISS</th>
<th>TVCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>$\theta_1$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>$\theta_1 \theta_2$</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>$\theta_1 \theta_3$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>$\theta_1 \theta_3 \theta_4$</td>
</tr>
</tbody>
</table>

Thus, $\theta_1$ is the reference value or typical value for a subject with no missing data having $x = 0$. $\theta_2$ is the proportional multiplier for subjects with no missing data having $x = 1$ and $\theta_3$ is the proportional multiplier for any subject with missing data. A similar model was used by Pitsiu et al. (2004) to model apparent oral clearance in an analysis where some subjects did not have creatinine clearance (CLcr) values

$$TVCL = \theta_1 + \theta_2 (1 - MISS) \text{CL}_{\text{CR}} + \theta_3 \text{MISS}. \tag{73}$$

Hence, subjects with missing data (MISS = 1) have the model

$$TVCL = \theta_1 + \theta_3 \tag{74}$$

and subjects without missing data (MISS = 0) have the model.

$$TVCL = \theta_1 + \theta_2 \text{CL}_{\text{CR}}. \tag{75}$$

When the proportion of missing data is small and the covariates are static, the dummy variable method seems to work reasonably well, but biased parameter estimates result when the percent of missing data becomes large (Jones 1996).

Mould et al. (2002), instead of using a conditional model to account for missing covariates, used a joint model that modeled the missing covariate simultaneously within the context of the overall pharmacokinetic model. In their analysis of topotecan pharmacokinetics in subjects with solid tumors, weight was known to be an important covariate but was missing in ~20% of the database. The function for males

$$\text{Weight} = \theta_1 \exp(\theta_2 [\text{BSA} - 1.73]) \exp(\theta_3 [\text{CL}_{\text{CR}} - 70]) \tag{76}$$

and for females

$$\text{Weight} = \theta_1 \exp(\theta_2 [\text{BSA} - 1.73]) \exp(\theta_3 [\text{CL}_{\text{CR}} - 70]) \theta_4 \tag{77}$$

was then used to model weight with the predicted value then fed into the model for clearance. In this manner, the joint function allowed replacement of the missing covariates with reasonable values and still accounted for correlations among the covariates. This approach has the advantage in that it can handle time-dependent covariates in theory. The down-side is that computations times are longer than other imputation approaches.

To illustrate these methods using a data set that does not contain time-dependent covariates, concentration–time data were simulated under a two-compartment model with intravenous administration after a single dose. Subjects were randomized to one of four dose groups: 25, 50, 100, or
200 mg. Two designs were studied: 75 subjects (~50:50 males:females) under a dense sampling schedule with samples collected at 0, 0.5, 1, 1.5, 2, 3, 4, 6, 8, 12, 18, 24, 36, 48, and 72 h after dosing and 300 subjects (~50:50 males:females) under a sparse sampling design where a single sample was randomly drawn from each of the following time intervals: 0.5 to 1 h, 1–3 h, 22–26 h, and 72–76 h. Hence, four samples were collected per subject under the sparse design. Clearance was simulated as a function of weight

\[
TVCL = 200(\text{Weight})^{0.7},
\]  

(78)

where weight was simulated from a log-normal distribution with a median of 55 kg and CV of 32% for females and a mean of 90 kg with a CV of 27% kg for males. A typical simulated dataset for the dense data had a median of 55 kg with a minimum of 44.6 kg and a maximum of 65.1 kg for females and a simulated median of 90.4 kg with a minimum of 76.3 kg and a maximum of 107.6 kg for males. In addition to weight, height, and age were simulated. Height was simulated to be log-normal in distribution with a median of 70 in. and 63 in. males and females, respectively, a CV of 5% in both sexes, and have a correlation with weight of 0.70. Age was simulated independent from height and weight from a normal distribution with a mean of 40 years old and a variance of 4 years. Central volume, intercompartmental clearance, and peripheral volume were all simulated to have a log-normal distribution with a median of 20 L, 12 L/h, and 125 L, respectively. BSV was set at 20% for each pharmacokinetic parameter with all pharmacokinetic parameters being independent of the others. Residual variability was simulated as a log-normal distribution with a CV of 7.5%.

Four data sets were then created where the weight for a subject was randomly set to missing (MCAR) with varying percentages from 0 to 40%. If a subject was chosen to have missing data, all weight information for that subject was set equal to missing. Each data set was analyzed using the following methods: complete case, the indicator method, the joint method, mean imputation, and multiple imputation. A two-compartment model was then fit to each data set using FOCE-I where clearance was modeled using

\[
TVCL = \theta_l (\text{Weight} / 70)^{\theta_2}.
\]  

(79)

For the joint method, weight was modeled using

\[
\text{Weight}^* = (\theta_6 + \theta_5 \text{Sex})(\text{Age} / 50)^{\theta_1} (\text{Height} / 60)^{\theta_8},
\]  

(80)

where Sex was coded as “0” for males and “1” for females. If a patient’s weight was missing, then Weight* was used in (79); otherwise, the observed weight was used. For the multiple imputation data sets, the MI procedure in SAS was used using monotone imputation with mean drug concentration, sex, age, and height as covariates to impute weight. Five imputation data sets were generated and each were analyzed. The mean parameter values from the multiple imputed data sets were reported as the final parameter values. Mean drug concentration (it could be argued that perhaps geometric mean concentration should have been used instead) was needed in the imputation model because only when the dependent variable is included in the imputation model are the resulting parameter estimates unbiased (Allison 2002). In fact, leaving out the dependent variable from an imputation model may result in spuriously small parameter estimates for those parameters associated with a covariate having missing data.

The results of the simulation are presented in Tables 5–9. No difference was observed between any of the imputation methods. Each of the imputation methods resulted in mean parameter estimates that could not be distinguished from the case where no data were missing. Even mean imputation, which is frequently criticized in the literature, did well. A tendency for the variability of the parameter estimates to increase as the percent missing data increased was observed across all methods, although little difference was observed in the variability of the parameter estimates across methods. If anything, multiple imputation performed the worst as the variability in some of the variance components was large compared to the other methods.

In summary, none of the methods were distinguishable from the others under any of the conditions used in the simulation. A key assumption to this conclusion was that the covariates were not time-dependent. Each covariate was a constant for an individual across all time points. Hence, the results from this simulation, while quite limited, suggest that under the right conditions, even when a large amount of information is missing, appropriate methods may be used to impute the data and recover the parameter estimates. These results also suggest that there may be little to no advantage of using one imputation method over another. But, these results need to be tempered with the fact that the covariates were static and that these results cannot be extrapolated to the case where the covariates are time-varying.

There is no one right way to handle missing data. Each data set must be approached on a case-by-case basis. What worked well in one situation may not work well in the next. What is important is being able to justify the method used to handle the missing data and make some evaluation of the impact in how the missing data were handled on the final model. The outcome of an analysis should not depend on the method used to handle the missing data. One way to evaluate the impact of the missing data treatment method is to conduct a sensitivity analysis or, if the proportion of missing data is small, to compare the model with the imputed data to the complete case model with the missing data excluded. Unfortunately, this section was far too brief to cover missing data in its entirety. Handling dropouts in a study was completely ignored. The reader is referred to Allison (2002) for a good overview of handling missing data in general, although he does not cover missing data in the longitudinal case to any great extent. Schafer (1997) and Little and Rubin (2002) are other more extensive, technical references.
### Table 5

Results from missing data analysis using complete case analysis

<table>
<thead>
<tr>
<th>Dense data sampling</th>
<th>True value</th>
<th>No data missing</th>
<th>10% missing</th>
<th>20% missing</th>
<th>40% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$ (mL/h)</td>
<td>–</td>
<td>3.97 (0.09)</td>
<td>3.97 (0.10)</td>
<td>3.96 (0.12)</td>
<td>3.97 (0.15)</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.7</td>
<td>0.70 (0.10)</td>
<td>0.72 (0.11)</td>
<td>0.69 (0.11)</td>
<td>0.67 (0.14)</td>
</tr>
<tr>
<td>V1 (L)</td>
<td>20</td>
<td>20.0 (0.5)</td>
<td>19.9 (0.6)</td>
<td>19.9 (0.6)</td>
<td>19.9 (0.9)</td>
</tr>
<tr>
<td>$Q$ (L/h)</td>
<td>12</td>
<td>12.1 (0.3)</td>
<td>12.0 (0.3)</td>
<td>12.0 (0.3)</td>
<td>12.0 (0.4)</td>
</tr>
<tr>
<td>V2 (L)</td>
<td>125</td>
<td>124.2 (2.4)</td>
<td>125 (2.1)</td>
<td>124 (2.6)</td>
<td>125 (3.5)</td>
</tr>
<tr>
<td>CV% (CL)</td>
<td>20</td>
<td>20.2 (1.7)</td>
<td>19.9 (1.8)</td>
<td>19.9 (2.0)</td>
<td>19.7 (2.0)</td>
</tr>
<tr>
<td>CV% (V1)</td>
<td>20</td>
<td>20.4 (1.8)</td>
<td>20.1 (1.7)</td>
<td>20.3 (1.9)</td>
<td>20.1 (2.1)</td>
</tr>
<tr>
<td>CV% ($Q$)</td>
<td>20</td>
<td>19.3 (1.5)</td>
<td>19.6 (1.3)</td>
<td>19.3 (1.6)</td>
<td>19.4 (2.2)</td>
</tr>
<tr>
<td>CV% (V2)</td>
<td>20</td>
<td>19.6 (1.2)</td>
<td>19.6 (1.6)</td>
<td>19.5 (1.8)</td>
<td>19.1 (2.5)</td>
</tr>
<tr>
<td>CV% ($\sigma^2$)</td>
<td>7.5</td>
<td>7.5 (0.2)</td>
<td>7.5 (0.2)</td>
<td>7.5 (0.2)</td>
<td>7.5 (0.3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sparse data sampling</th>
<th>True value</th>
<th>No data missing</th>
<th>10% missing</th>
<th>20% missing</th>
<th>40% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$ (mL/h)</td>
<td>–</td>
<td>3.95 (0.04)</td>
<td>3.96 (0.04)</td>
<td>3.96 (0.04)</td>
<td>3.96 (0.05)</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.7</td>
<td>0.70 (0.05)</td>
<td>0.70 (0.06)</td>
<td>0.70 (0.06)</td>
<td>0.69 (0.06)</td>
</tr>
<tr>
<td>V1 (L)</td>
<td>20</td>
<td>20.2 (0.3)</td>
<td>20.2 (0.2)</td>
<td>20.2 (0.3)</td>
<td>20.1 (0.4)</td>
</tr>
<tr>
<td>$Q$ (L/h)</td>
<td>12</td>
<td>12.1 (0.2)</td>
<td>12.1 (0.2)</td>
<td>12.1 (0.2)</td>
<td>12.1 (0.2)</td>
</tr>
<tr>
<td>V2 (L)</td>
<td>125</td>
<td>126 (2.1)</td>
<td>126 (2.1)</td>
<td>126 (2.1)</td>
<td>126 (2.2)</td>
</tr>
<tr>
<td>CV% (CL)</td>
<td>20</td>
<td>19.9 (2.0)</td>
<td>20.1 (1.9)</td>
<td>19.7 (2.1)</td>
<td>19.6 (1.9)</td>
</tr>
<tr>
<td>CV% (V1)</td>
<td>20</td>
<td>20.1 (1.4)</td>
<td>20.0 (1.5)</td>
<td>20.0 (1.5)</td>
<td>20.0 (1.8)</td>
</tr>
<tr>
<td>CV% ($Q$)</td>
<td>20</td>
<td>19.6 (1.2)</td>
<td>19.5 (1.6)</td>
<td>19.5 (1.6)</td>
<td>19.1 (1.7)</td>
</tr>
<tr>
<td>CV% ($\sigma^2$)</td>
<td>7.5</td>
<td>7.2 (0.8)</td>
<td>7.1 (0.9)</td>
<td>7.1 (0.9)</td>
<td>7.1 (0.9)</td>
</tr>
</tbody>
</table>

### Table 6

Results from missing data analysis using the indicator method

<table>
<thead>
<tr>
<th>Dense data sampling</th>
<th>True value</th>
<th>No data missing</th>
<th>10% missing</th>
<th>20% missing</th>
<th>40% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$ (mL/h)</td>
<td>–</td>
<td>3.97 (0.09)</td>
<td>3.96 (0.11)</td>
<td>3.97 (0.12)</td>
<td>3.97 (0.15)</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.7</td>
<td>0.70 (0.10)</td>
<td>0.71 (0.11)</td>
<td>0.70 (0.12)</td>
<td>0.67 (0.14)</td>
</tr>
<tr>
<td>V1 (L)</td>
<td>20</td>
<td>20.0 (0.5)</td>
<td>20.0 (0.5)</td>
<td>20.0 (0.5)</td>
<td>20.0 (0.5)</td>
</tr>
<tr>
<td>$Q$ (L/h)</td>
<td>12</td>
<td>12.0 (0.3)</td>
<td>12.0 (0.3)</td>
<td>12.0 (0.3)</td>
<td>12.0 (0.3)</td>
</tr>
<tr>
<td>V2 (L)</td>
<td>125</td>
<td>124 (2.4)</td>
<td>124 (2.3)</td>
<td>124 (2.4)</td>
<td>124 (2.4)</td>
</tr>
<tr>
<td>CV% (CL)</td>
<td>20</td>
<td>20.2 (1.7)</td>
<td>20.6 (1.8)</td>
<td>21.6 (1.6)</td>
<td>23.3 (2.0)</td>
</tr>
<tr>
<td>CV% (V1)</td>
<td>20</td>
<td>20.4 (1.8)</td>
<td>20.3 (1.8)</td>
<td>20.3 (1.8)</td>
<td>20.3 (1.8)</td>
</tr>
<tr>
<td>CV% ($Q$)</td>
<td>20</td>
<td>19.3 (1.5)</td>
<td>19.4 (1.3)</td>
<td>19.4 (1.3)</td>
<td>19.3 (1.3)</td>
</tr>
<tr>
<td>CV% (V2)</td>
<td>20</td>
<td>19.6 (1.2)</td>
<td>19.5 (1.6)</td>
<td>19.5 (1.6)</td>
<td>19.8 (1.5)</td>
</tr>
<tr>
<td>CV% ($\sigma^2$)</td>
<td>7.5</td>
<td>7.5 (0.2)</td>
<td>7.5 (0.2)</td>
<td>7.5 (0.2)</td>
<td>7.5 (0.2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sparse data sampling</th>
<th>True value</th>
<th>No data missing</th>
<th>10% missing</th>
<th>20% missing</th>
<th>40% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$ (mL/h)</td>
<td>–</td>
<td>3.95 (0.04)</td>
<td>3.96 (0.04)</td>
<td>3.96 (0.04)</td>
<td>3.96 (0.05)</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.7</td>
<td>0.70 (0.05)</td>
<td>0.71 (0.06)</td>
<td>0.70 (0.06)</td>
<td>0.69 (0.06)</td>
</tr>
<tr>
<td>V1 (L)</td>
<td>20</td>
<td>20.2 (0.3)</td>
<td>20.2 (0.3)</td>
<td>20.2 (0.3)</td>
<td>20.2 (0.3)</td>
</tr>
<tr>
<td>$Q$ (L/h)</td>
<td>12</td>
<td>12.1 (0.2)</td>
<td>12.1 (0.2)</td>
<td>12.1 (0.2)</td>
<td>12.1 (0.2)</td>
</tr>
<tr>
<td>V2 (L)</td>
<td>125</td>
<td>126 (2.1)</td>
<td>126 (2.1)</td>
<td>126 (2.1)</td>
<td>126 (2.1)</td>
</tr>
<tr>
<td>CV% (CL)</td>
<td>20</td>
<td>20.0 (0.7)</td>
<td>20.7 (0.8)</td>
<td>21.6 (0.9)</td>
<td>23.3 (1.1)</td>
</tr>
<tr>
<td>CV% (V1)</td>
<td>20</td>
<td>19.9 (2.0)</td>
<td>19.8 (2.0)</td>
<td>19.8 (2.0)</td>
<td>19.8 (2.1)</td>
</tr>
<tr>
<td>CV% ($Q$)</td>
<td>20</td>
<td>20.0 (1.3)</td>
<td>20.0 (1.3)</td>
<td>20.0 (1.3)</td>
<td>20.0 (1.3)</td>
</tr>
<tr>
<td>CV% (V2)</td>
<td>20</td>
<td>20.0 (1.5)</td>
<td>20.0 (1.5)</td>
<td>20.0 (1.5)</td>
<td>20.0 (1.5)</td>
</tr>
<tr>
<td>CV% ($\sigma^2$)</td>
<td>7.5</td>
<td>7.2 (0.8)</td>
<td>7.2 (0.8)</td>
<td>7.2 (0.8)</td>
<td>7.2 (0.8)</td>
</tr>
</tbody>
</table>
### Table 7
Results from missing data analysis using the joint model

<table>
<thead>
<tr>
<th>Dense data sampling</th>
<th>True value</th>
<th>No data missing</th>
<th>10% missing</th>
<th>20% missing</th>
<th>40% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$ (mL/h)</td>
<td>–</td>
<td>3.97 (0.09)</td>
<td>3.99 (0.09)</td>
<td>3.97 (0.13)</td>
<td>3.99 (0.14)</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.7</td>
<td>0.7 (0.10)</td>
<td>0.69 (0.19)</td>
<td>0.7 (0.13)</td>
<td>0.66 (0.20)</td>
</tr>
<tr>
<td>V1 (L)</td>
<td>20</td>
<td>20.0 (0.5)</td>
<td>20.1 (0.5)</td>
<td>20.0 (0.6)</td>
<td>20.0 (0.5)</td>
</tr>
<tr>
<td>Q (L/h)</td>
<td>12</td>
<td>12.0 (0.3)</td>
<td>12.0 (0.3)</td>
<td>12.0 (0.3)</td>
<td>12.0 (0.3)</td>
</tr>
<tr>
<td>V2 (L)</td>
<td>125</td>
<td>124 (2.4)</td>
<td>124 (2.4)</td>
<td>124 (2.6)</td>
<td>124 (2.5)</td>
</tr>
<tr>
<td>CV% (CL)</td>
<td>20</td>
<td>20.2 (1.7)</td>
<td>19.9 (2.4)</td>
<td>19.8 (1.7)</td>
<td>20.3 (2.0)</td>
</tr>
<tr>
<td>CV% (V1)</td>
<td>20</td>
<td>20.4 (1.8)</td>
<td>20.3 (1.7)</td>
<td>20.5 (1.8)</td>
<td>20.4 (1.8)</td>
</tr>
<tr>
<td>CV% (Q)</td>
<td>20</td>
<td>19.3 (1.5)</td>
<td>19.4 (1.2)</td>
<td>19.5 (1.3)</td>
<td>20.0 (1.3)</td>
</tr>
<tr>
<td>CV% (V2)</td>
<td>20</td>
<td>19.6 (1.2)</td>
<td>19.4 (1.7)</td>
<td>19.6 (1.3)</td>
<td>19.6 (1.3)</td>
</tr>
<tr>
<td>CV% ($\sigma$)</td>
<td>7.5</td>
<td>7.5 (0.2)</td>
<td>7.5 (0.2)</td>
<td>7.5 (0.2)</td>
<td>7.5 (0.2)</td>
</tr>
</tbody>
</table>

Table 8
Results from missing data analysis using mean imputation

<table>
<thead>
<tr>
<th>Sparse data sampling</th>
<th>True value</th>
<th>No data missing</th>
<th>10% missing</th>
<th>20% missing</th>
<th>40% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$ (mL/h)</td>
<td>–</td>
<td>3.95 (0.04)</td>
<td>3.96 (0.04)</td>
<td>3.96 (0.04)</td>
<td>3.97 (0.04)</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.7</td>
<td>0.7 (0.05)</td>
<td>0.7 (0.06)</td>
<td>0.7 (0.06)</td>
<td>0.69 (0.06)</td>
</tr>
<tr>
<td>V1 (L)</td>
<td>20</td>
<td>20.2 (0.3)</td>
<td>20.2 (0.2)</td>
<td>20.2 (0.3)</td>
<td>20.2 (0.3)</td>
</tr>
<tr>
<td>Q (L/h)</td>
<td>12</td>
<td>12.1 (0.2)</td>
<td>12.1 (0.2)</td>
<td>12.1 (0.2)</td>
<td>12.1 (0.2)</td>
</tr>
<tr>
<td>V2 (L)</td>
<td>125</td>
<td>126 (2.1)</td>
<td>126 (2.1)</td>
<td>126 (2.1)</td>
<td>126 (2.2)</td>
</tr>
<tr>
<td>CV% (CL)</td>
<td>20</td>
<td>20.0 (0.7)</td>
<td>19.8 (0.7)</td>
<td>20.0 (0.7)</td>
<td>20.2 (0.7)</td>
</tr>
<tr>
<td>CV% (V1)</td>
<td>20</td>
<td>19.9 (2.0)</td>
<td>20.0 (2.0)</td>
<td>19.9 (2.0)</td>
<td>19.9 (2.0)</td>
</tr>
<tr>
<td>CV% (Q)</td>
<td>20</td>
<td>20.0 (1.3)</td>
<td>20.0 (1.3)</td>
<td>20.0 (1.3)</td>
<td>20.0 (1.3)</td>
</tr>
<tr>
<td>CV% (V2)</td>
<td>20</td>
<td>20.0 (1.5)</td>
<td>20.1 (1.4)</td>
<td>20.0 (1.5)</td>
<td>19.9 (1.5)</td>
</tr>
<tr>
<td>CV% ($\sigma$)</td>
<td>7.5</td>
<td>7.2 (0.8)</td>
<td>7.2 (0.8)</td>
<td>7.2 (0.8)</td>
<td>7.3 (0.8)</td>
</tr>
</tbody>
</table>
Internal Validity Checks and Data Clean-Up

One of the most time consuming tasks in any PopPK analysis is not necessarily model development but developing the datasets used for model building and validation. It is imperative that these datasets are formatted correctly or any analysis will be wasted and possibly incorrect. Prior to any modeling the database should be scrutinized for possible errors. The dates and time of administered doses, reported concentrations, demographic information, and other subject information should be examined for missing data or suspected errors. It is not uncommon for PopPK databases to be merged across many clinical studies which are themselves pooled across many different study centers. For laboratory values in particular, care should be taken that all the variables have the same units. It is not uncommon for sites outside the USA to use the metric system, while sites in the USA do not. Hence, all covariates should be checked for unit consistency. A quick test for consistency is to examine the frequency distribution of the covariate. Multimodal covariate is usually indicative of a units problem. If any data are missing, consistency in handling of missing values or concentration data below the LLOQ should be verified. Details on how these errors and inconsistencies will be corrected should be detailed in the DAP. Any exceptions to the DAP should be documented. Also, any changes or manipulations to the database should be clearly documented. Once all these data set issues have been resolved, model development can proceed.

Problems and Errors

Abnormal terminations during the optimization process or nonconvergence with nonlinear mixed effects models are usually due to an overparameterized covariance matrix or poor initial estimates and can be overcome by a few approaches (Pinheiro and Bates 1994). One approach is to treat the variance–covariance matrix as diagonal, treating off-diagonal covariances as zero. By examining the variance components, one or more may be so small so as to be removed from the model. Another approach is to force convergence by setting the maximal number of iterations to some value and then examine the covariance matrix for elements that can be removed. Sometimes, a covariance term may be small enough that the term can be removed. Pinheiro and Bates (1994) suggest examining the eigenvalues of the covariance matrix for overparameterization. If the smallest eigenvalue is small relative to the others or is near zero, this is indicative of overparameterization. Alternatively, one can compute the condition number of the covariance matrix and see how large it is. A large condition number is indicative of an unstable model and should be simplified. If a covariance matrix indicates that some covariance terms are needed but some are not then perhaps the covariance matrix can be reformulated to a banded matrix where some off-diagonal elements are fixed equal to zero. If the problem is not the covariance matrix, then possibly the initial estimates are poor. Check the magnitude of the starting values to make sure they are the correct scale, e.g., clearance should be in mL/h but the initial estimate is in
Table 10
Common NONMEM error messages and some solutions

<table>
<thead>
<tr>
<th>Example error message</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MINIMIZATION TERMINATED</strong> DUE TO PROXIMITY OF LAST ITERATION EST. TO A VALUE AT WHICH THE OBJ. FUNC. IS INFINITE (ERROR=136, 138)</td>
<td>• This message is telling you that at the point of termination an infinite objective function was encountered, which may have been due to a possible zero concentration in one of the observations. Check and remove. Or add a very small additive error term to the error model, e.g., ( Y = F \times \exp(\text{EPS}(1)) + \text{EPS}(2) ), where ( \text{EPS}(2) ) is fixed to something small (&lt;0.001). Alternatively, try \text{METHOD} = \text{HYBRID}.</td>
</tr>
<tr>
<td><strong>DUE TO PROXIMITY OF NEXT ITERATION EST. TO A VALUE AT WHICH THE OBJ. FUNC. IS INFINITE (ERROR=137)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>DUE TO INFINITE INITIAL VALUE OF OBJ. FUNC. AT INITIAL ESTIMATE (ERROR=135)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>MINIMIZATION TERMINATED</strong> PARAMETER ESTIMATE IS NEAR ITS BOUNDARY THIS MUST BE ADDRESSED BEFORE THE COVARIANCE STEP CAN BE IMPLEMENTS (ERROR=139)**</td>
<td>• This message is telling you that one of the model parameters is at a boundary, usually a zero bound. The boundary needs to be relaxed in the $\text{THETA} ) block or the parameter needs to be reparameterized to relax the boundary</td>
</tr>
<tr>
<td><strong>Floating overflow</strong></td>
<td>• This indicates that a division by zero has occurred. Check the control stream for functions that include a division by a variable term and then examine that variable in the data set for any values set equal to zero. It has been reported that when everything else has failed, sometimes a new Fortran compiler can solve the problem</td>
</tr>
<tr>
<td><strong>ETABAR test statistic reported by NONMEM not equal to zero</strong></td>
<td>• If the distribution of the random effects is independent and normally distributed, the model parameter estimates are maximum likelihood estimates. A nonzero ( \text{ETABAR} ) test, which is a ( p )-value, indicates that the mean of a random effect (the ETAs) is not zero. This may be due to a covariate not included in the model, shrinkage of the ETAs (if equal shrinkage occurs on both sides of the distribution then this may not be an issue), incorrect distribution assumptions regarding the ETAs, e.g., a log-normal distribution may be more appropriate, or if a very large number of subjects are in the dataset, the ( p )-value may be an artifact of the sample size. In all cases, look at the value of ( \text{ETABAR} ) to see how far removed from zero the value is</td>
</tr>
</tbody>
</table>

(Continued)
Table 10
Common NONMEM error messages and some solutions

<table>
<thead>
<tr>
<th>Example error message</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCCURS DURING SEARCH FOR ETA AT A NONZERO VALUE OF ETA K21, OR K31 IS TOO CLOSE TO AN EIGENVALUE 0</td>
<td>The error message is the result of running the POSTHOC step. Look at the individual where the error occurs and see if there is something unusual about that person’s data. Try using the NOABORT option on the $COV statement. Try changing the starting values</td>
</tr>
<tr>
<td>PROGRAM TERMINATED BY OBJ ERROR IN CELS WITH INDIVIDUAL 5 ID=.50000000E+01 WEIGHTED SUM OF “SQUARED” INDIVIDUAL RESIDUALS IS INFINITE MESSAGE ISSUED FROM ESTIMATION STEP AT INITIAL OBJ. FUNCTION EVALUATION</td>
<td>This indicates that the residual sum of squares for the individual is near zero. Check the initial value of the residual error and see if it is small. If so, increase its value. Alternatively, try fitting the logarithms of the concentration with an additive residual error model</td>
</tr>
<tr>
<td>Number of function evaluations exceeded MAXEVAL (ERROR=131)</td>
<td>In the $COV step change the MAXEVAL= option to a larger value or restart the model at the current, last set of parameter estimates</td>
</tr>
<tr>
<td>288 SIZE OF NMPRD4 EXCEEDED; LNP4 IS TOO SMALL IN NM-TRAN AND NONMEM</td>
<td>Your model may have exceeded the maximum number of fixed effects and variance components. You may need to increase TSIZE and NSIZE within NONMEM</td>
</tr>
<tr>
<td>(WARNING 2) NM-TRAN INFERS THAT THE DATA ARE POPULATION.</td>
<td>This indicates there is a nonnumeric field in the NONMEM dataset. NONMEM only accepts numeric data and will not accept character data. This error can also occur when the first character of a new line in the dataset is meant to indicate “skip this line” and IGNORE=c (where c is the ignoring character, usually C) is not included in the $DATA line in the control stream</td>
</tr>
<tr>
<td>(DATA ERROR) RECORD 1, DATA ITEM 1, CONTENTS: ID ITEM IS NOT A NUMBER.</td>
<td></td>
</tr>
<tr>
<td>MINIMIZATION SUCCESSFUL HOWEVER, PROBLEMS OCCURRED WITH THE MINIMIZATION. REGARD THE RESULTS OF THE ESTIMATION STEP CAREFULLY, AND ACCEPT THEM ONLY AFTER CHECKING THAT THE COVARIANCE STEP PRODUCES REASONABLE OUTPUT</td>
<td>This message relates to the covariance matrix and may happen when the gradient is too large. Look for parameters with small standard errors and a large final gradient. The covariance matrix may be overparameterized but in this case it may be that the variance estimate is not necessarily close to zero. Try rescaling your model parameter to have similar scale. Try using the UNCONDITIONAL option on the $COV step or using MATRIX=S option</td>
</tr>
<tr>
<td>Example error message</td>
<td>Solution</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>TOT. NO. OF PRED-DEFINED RECS IN BUFFER 6 IS LESS THAN NO. OF DATA RECS WITH SOME INDIVIDUAL</td>
<td>• Some ID has more data records than defined by NSIZES. Need to increase NSIZES and recompile NONMEM</td>
</tr>
<tr>
<td><strong>R</strong> MATRIX ALGORTHMICALLY SINGULAR AND ALGORTHMICALLY NON-POSITIVE-SEMIDEFINITE COVARIANCE STEP ABORTED</td>
<td>• This message is not necessarily an error message but a warning. This error is usually due to a model that is overparameterized. Look for parameters, particularly variance components, that are very small, e.g., (&lt;1 \times 10^{-6}), and remove them from the model. What you are trying to do here is simplify the model</td>
</tr>
<tr>
<td>OR</td>
<td></td>
</tr>
<tr>
<td><strong>S</strong> MATRIX ALGORTHMICALLY SINGULAR AND ALGORTHMICALLY NON-POSITIVE-SEMIDEFINITE COVARIANCE STEP ABORTED</td>
<td></td>
</tr>
<tr>
<td>MINIMIZATION TERMINATED DUE TO ROUNING ERRORS (ERROR=134)</td>
<td>• Computers are able to retain numbers within their memory only to a certain number of digits. During the computing of a mathematical operation, such as matrix inversion, digits in a number may be lost. This message indicates that a sufficient number of digits was lost resulting in termination of the optimization. One solution is to rerun the model with a new set of starting values and change the number of significant digits to higher than the current value. In this example SIGDIGITS=3 in the $COV step. One solution would be to set SIGDIGITS=4. The starting values should then be set equal to the final estimates at the point the algorithm terminated. This process may have to be repeated many times. A second solution is to decrease the number of significant digits, such as to SIGDIGITS=2 in this example. If the rounding error is a variance component then this is usually an acceptable solution. A third solution is to simplify the model because the model is overparameterized. A fourth solution is to ignore the error or to remove the $COV step. A fifth step is to try the SLOW option on the $COV step. A fourth solution is to reorder the individuals in the dataset. Sale (personal communication) indicates that rounding errors may sometimes be an effect of a floating-point arithmetic computation and that by changing the order of the patients, not the observations within a patient, that some rounding errors could be avoided or at least solved.</td>
</tr>
</tbody>
</table>
L/h. Alternatively, a quick user-defined grid search can be done to find starting values by setting the maximum number of evaluations equal to zero and looking to see how the objective function value changes. A number of other errors may occur during routine use with NONMEM which are listed in Table 10 along with some solutions.

### Consistency of Model Parameter Estimates Across Computer Platforms

NONMEM is provided on UNIX, DOS, and Macintosh platforms. Although no formal studies have been done, there is some data to suggest that different platforms may result in different population estimates. Frame et al. (2001) reported that in a population analysis, a particular model could not converge on the DOS platform due to either rounding errors or aborted covariance step, but could do so on the UNIX platform. Curiously, the parameter estimates for both platforms under the final converged model (UNIX) and model at this time of failure (DOS) were almost identical. Bonate (2002), with the help of many contributors, compared the consistency across users of five different models in NONMEM-reported parameter estimates and their standard errors. All users used NONMEM V on a personal (31/38, 81%), Unix (6/38, 16%), or Macintosh (1/38, 3%) computer. Ten different compilers were tested. In those models that optimized without errors, the estimates of the fixed effects and variance components were 100% consistent across users, although there were some small differences in the estimates of the standard error of the parameter estimates. Different compilers produced small differences in the estimates of the standard errors. The standard errors were not consistent even with users running the same compiler version.

With one data set, 27 of 38 testers could not minimize the run successfully without abnormal terminations: 5 due to infinite objective function, 12 due to maximum function evaluations exceeded, and 10 due to rounding errors. Only 1 user, who used a UNIX computer running HP-UX Fortran 77, reported successful minimization without errors. Interestingly, two other users running the same hardware--software combination reported termination due to maximum number of function evaluations exceeded or termination due to rounding errors. In the 11 testers who could successfully minimize the data set, 10 of 11 reported R-matrix singularity. These results suggest that models that fail to minimize successfully on one platform and compiler may not fail on another, and vice-versa. These results are highly suggestive that NONMEM results are machine specific, although the differences across machines may be small enough to be ignored. No such other studies have been done for other software systems, such as the NLMIXED procedure in SAS or with the NLME function in S-PLUS, and no studies have been done with later versions of NONMEM. It does seem likely, however, that the compiler differences reported with earlier versions of NONMEM will continue with later versions. Personal experience indicates this to be true as I have seen models in NONMEM 7 have slightly different standard errors on a personal computer compared to a UNIX workstation.

### Mu-Referencing

With the release of NONMEM 7 and the accompanying suite of EM estimation algorithms, in order to be able to efficiently use these algorithms, model parameters must be mu-referenced thereby allowing NONMEM to effectively link an individual parameter estimate (called PHI) to a THETA and ETA. Failure to do mu-referencing for more than 2 THETAs will not result in valid parameter estimates but will result in inefficient, slower parameter estimation, and possible nonconvergence.

In its most basic form, suppose CL is parameterized in NONMEM as

\[ CL = THETA(1) \times ETA(1) \]

In mu-referencing, CL would be reparameterized as

\[ MU_1 = THETA(1) \\
CL = EXP(MU_1 + ETA(1)) \]

These statements are equivalent because of the rules of exponentiation, although the values of THETA(1) are different. If it was of interest to maintain parameters on the original scale, an equivalent parameterization would be

\[ MU_1 = LOG(THETA(1)) \\
CL = EXP(MU_1 + ETA(1)) \]

In this manner, MU is modeled additively with ETA, and that is the key to mu-referencing, the reference must be additive in nature and the index of MU_i and ETA(i) must match, e.g., MU_1 and ETA(1), MU_2 and ETA(2), etc., so that there is a linear, one-to-one relationship between MU and ETA.

As a further example, if CL was modeled as a function of weight

\[ CL = THETA(1) \times (WGT/80)^{THETA(2)} \times EXP(ETA(1)) \]

Then the mu-referenced form would be

\[ MU_1 = LOG(THETA(1)) \times (WGT/80)^{THETA(2)} \\
CL = EXP(MU_1 + ETA(1)) \]

Or equivalently

\[ TVCL = THETA(1) \times (WGT/80)^{THETA(2)} \\
MU_1 = LOG(TVCL) \\
CL = EXP(MU_1 + ETA(1)) \]

Additional efficiencies are gained if MU can be written in an additive form since this allows Gibbs sampling to be implemented instead of Metropolis–Hastings sampling, the former being faster than the latter. For example, the previous example could be rewritten as

\[ MU_1 = LOG(THETA(1)) + THETA(2) \times LOG(WGT/80) \\
CL = EXP(MU_1 + ETA(1)) \]

In which the entire statement for MU_1 is linear. The trick to mu-referencing is that MU and ETA must be linear and that the model for the parameter of interest does not change. In all three instances above, CL is equivalently defined.

There are some other rules regarding mu-referencing. First, if a THETA is referenced to a MU, that THETA cannot be referenced later in the model code. For instance, if SEX were coded as “0” for males and “1” for females, then is incorrect because THETA(1) is referenced again
after it has been mu-referenced. Note that the linebreaks shown above are not allowed in NONMEM but are used herein for formatting. The correct form would be

Second, model parameters without an ETA cannot be mu-referenced. For example, if V3 = THETA(4), then all subjects have the same V3 and mu-referencing cannot be done. However, it has been recommended that if the model allows, it is better to assign an ETA to the parameter and fix it to a very small value. For example, V3 could be mu-referenced as

\[
\text{MU}_3 = \log(\text{THETA}(4)) \\
V3 = \exp(\text{MU}_3 + \text{ETA}(4)) \\
\text{SOME}_\text{G} = 0.0001 \text{ FIX}
\]

Third, covariates that vary with observation, dose, or time cannot be mu-referenced. For example, if age, weight, or some laboratory value varies over time, these cannot be mu-referenced. There are, however, special cases where the mu-reference can be tricked by rearranging the model. Suppose V was proportional to weight (WGT), as is the case with aminoglycoside antibiotics, then a reasonable model would be

\[
V = (\text{THETA}(2) \times (\text{WGT}/80)) \times \exp(\text{ETA}(1))
\]

The corresponding mu-referenced model would be

\[
\text{MU}_2 = \log(\text{THETA}(2)) \\
V = (\text{WGT}/80) \times \exp(\text{MU}_2 + \text{ETA}(2))
\]

In this case, the covariate is pulled outside the mu-reference. Lastly, mu-referencing cannot be done in a SERROR block, should not be used with IF-THEN statements, should be done as early as possible in the SPK or SPRED block, and must be defined prior to any verbatim code.

NONMEM also has an option in the $EST statement, called MUM, that allows the user to specify whether a mu-reference will be used with a THETA. MUM can take values of M to indicate mu-referencing, N for not mu-referenced, and D, which is the default option and in NONMEM 7 is currently equivalent to M. For example, suppose there are 5 THETAs and the last THETA will not be mu-referenced, the $EST statement

\[
\text{SEST... MUM=M MMMN}
\]

could be used. With mixture models, the MUM option must be used. So, if the $MIX statement was

\[
\text{SMIX} \\
\text{N}\text{SPOP} = 2 \\
(P(1) = \text{THETA}(5) \\
P(2) = 1 - P(5)
\]

The $EST statement would be as specified previously.

Sometimes, the residual variance model is defined as using a THETA instead of a SIGMA. For example, the model

\[
Y = F \times \exp(\text{EPS}(1)) \\
\text{SSSIGMA} = 0.1
\]
is equivalent to

\[
Y = F \times \exp(\text{THETA}(3)) \\
\text{STTHETA} = 0.1 ; \text{THETA}(3) \\
\text{SSSIGMA} = \text{FIXED}
\]

These THETAs are referred to as “sigma-like” thetas and cannot be mu-referenced, but if these are specified in the GRD option of the $EST statement with the value “S,” the speed of the algorithm can be improved. GRD can take values of D for let NONMEM decide, S for sigma-like, G for Gibbs sampling, and N for Metropolis–Hastings sampling. G and N are only used with BAYES estimation. Hence, GRD would be specified as GRD = DDS in this example.

In summary, mu-referencing associates MU_x with the ETA(x) in a linear, additive manner to allow more efficient estimation with EM algorithms. Mu-referencing has no effect with the maximum likelihood estimation algorithms, like FOCE.

**Regulatory Review of Population Analyses**

Population analyses are usually targeted to one of three audiences: the medical literature, internal review, or regulatory reviewers. When the objective of an analysis is a publication, model development is usually not as stringent because quite often most medical journals in which population analyses are published are limited in the amount of page space that can be devoted to methods and so often analyses omit the details on method development and focus on the “big picture.” The same can be said of analyses conducted for internal review but for different reasons.

Often the outcome of an internal analysis is a presentation to project team members or management and a report may never even be written.

In contrast, the level of detail in a report can be overwhelming when the audience is regulatory authorities and in this section, attention will focus on the FDA. It is common for a PopPK report, not even including appendices, to total hundreds of pages with the number of tables and figures in the mid-double digits. Already discussed has been a sample report template in this setting (see Table 2). What is important with a PopPK report for regulatory authorities is that the report has a clear flow and logic for how model development proceeded and that the conclusions were supported by the analysis. Further, FDA reviewers have stated that a detailed table of contents with many report body headings are very useful for them in their review. It is always useful to write with the reviewer in mind. To often we write to a level of obfuscation (and hopefully you do not think that with this book). Writing with clear messages and issues with logical analysis and interpretation will aid in the review of any report. An internal government review at the FDA showed that 70% of responding reviewers reported that poor document quality slowed their review. FDA reviewers are just like you – they are busy – so make your reports easy for them to review.

Reviewers at the FDA now take a question-based approach in reviewing an NDA (Lesko and Williams 1999). Reviewers develop particular questions and then review the NDA as a means of answering those questions. The particular questions raised might include what are the general attributes of the drug, what are the general clinical pharmacology characteristics of the drug, what intrinsic and extrinsic factors affect drug exposure, what are the particular biopharmaceutical characteristics of the drug, is
there a need for all doses the manufacturer will make (e.g.,
the sponsor may market the drug in 5, 10, and 20 mg tablets
and the FDA may question whether a 5 mg tablet is
necessary), and were the bioanalytical methods sufficiently
accurate and precise in quantifying the drug and its
metabolites? Many of these questions are answered by a
single reviewer, but the review of the population analysis is
often done by a different reviewer, one with specific
expertise in population analyses. At the FDA this is often
the purview of the pharmacometrics group (http://www.fda.gov/AboutFDA/CentersOffices/CDER/ucm167032.htm,
accessed October 2010).

The FDA routinely publishes on the internet their
reviews of sponsor’s population analyses in support of a
New Drug Application (NDA). It is very informative to
read some of these to understand how regulatory authorities
review these reports. It is not uncommon to see back-seat
modeling from regulatory reviewers as they examine the
validity of the reported final model. One of the things a
reviewer is tasked to do is to understand and test the limits
of the author’s claims. The reviewer then may try to
develop better, alternative models that more precisely
estimate maximal concentrations. Failing that the reviewer
may then examine what impact this underprediction has on
the conclusions of the analysis or a pharmacokinetic–
pharmacodynamic analysis of the drug.

Unfortunately, there is no standardized policy or
procedures guide for reviewing PopPK analyses so
these reviews tend to be reviewer-specific. However,
pharmacometric reviewers tend to focus on the following
areas: data checking (how was the database developed and
examined for outliers or merge errors, what did the database
consist of, how many subjects, how many observations total
and per subject overall, did the database include sufficient
numbers of subjects in subpopulations to detect any
meaningful relationships with exposure), structural model
development (how was the structural model developed,
what estimation algorithms were used, what random effects
were added and why), covariate models (how were
covariates chosen for evaluation, reasonableness of the
covariates, any particular obvious covariates not examined,
how were covariate models examined, what estimation
algorithm was used, which model types were examined, e.g.,
linear or power, and reasonableness of the final covariate
model), validation (what validation techniques were used
and how reasonable were the results and conclusions),
software (which software was used), and does the analysis
contribute to the overall understanding of the drug. Given
then the reasonableness of a PopPK model the impact of the
model on dosing must be examined. Does the sponsor’s
dosing recommendations concur with the results of the
population analysis? Should some particular groups receive
dose reductions, e.g., if creatinine clearance is identified as
a covariate should patients with renal impairment be dosed
differently than patients with normal renal function?

Every effort should be made to make reports to
regulatory agencies as clear and concise as possible
with the report written to support the product label and
further help in understanding the pharmacology and
pharmacokinetics of the drug. In order to facilitate review
of these reports, sponsors should provide to the regulatory
authorities electronic copies of all data sets and control
streams in an ASCII format. A hard copy of the first page
of each data set should be included in the report with each
individual data column identified. Any SAS programs or
programs used to prepare the data sets should be made
available. Whatever would be useful for a reviewer to
repeat the modeling process from beginning to end should
be made available in an easily retrievable format.

Summary
Population pharmacokinetic analyses are fraught with
problems. This chapter was meant to explore these issues
and present real-world problems and solutions to the reader.
Some topics, like missing data, could not be covered in
their entirety. Indeed, entire books have been written on
some of the topics just briefly covered herein. Obviously
the reader is referred to these texts for greater detail. With
the topics presented in the last chapter on theory and this
chapter on practical issues, the introduction of real-world
PopPK analyses will be made in the next chapter.

Recommended Reading
Aarons L, Balant LP, Mentre F, Morselli PL, Rowland M,
Steimer J-L, Vozeh S. Practical experience and
issues in designing and performing population
pharmacokinetic/pharmacodynamic studies. European


European Medicines Agency, Committee for Medicinal
Products for Human Use. Guideline on Reporting the
Results of Population Pharmacokinetic Analyses,
18599006enfin.pdf).

Food and Drug Administration. Guidance for Industry:

Kowalski K, Hutmacher K. Design evaluation for a
population pharmacokinetic study using clinical trial
simulations: a case study. Statistics in Medicine 2001;
20: 75-91.

Lee PID. Design and power of a population pharmacokinetic

Wahlby U, Jonsson EN, Karlsson MO. Assessment of
actual significance levels for covariate effects in
NONMEM. Journal of Pharmacokinetics and

Wright PMC. Population based pharmacokinetic analysis:
Why do we need it; what is it; and what has it told us
about anesthesia. British Journal of Anesthesia 1998;
80: 488-501.
References


Beal SL. Ways to fit a PK model with some data below the quantification limit. *Journal of Pharmacokinetics and Pharmacodynamics* 2001; 28: 481-504.


DuBois D and DuBois EF. A formula to estimate the approximate surface area if height and weight be known. *Archives of Internal Medicine* 1916; 17: 863-871.


Kaila N, Straka RJ, and Brundage RC. Mixture models and subpopulation classification: a pharmacokinetic simulation study and application to metoprolol CYP2D6 phenotype. *Journal of Pharmacokinetics and Pharmacodynamics* 2006; 34: 141-156.


LaVeist TA. Beyond dummy variables and sample selection: what health services researchers ought to know about race as a variable. *Human Services Research* 1994; 29: 1-16.


