Bootstrapping for Pharmacokinetic Models: Visualization of Predictive and Parameter Uncertainty

C. Anthony Hunt,[1,4] Geof H. Givens,[2] and Serge Guzy[3]

Received November 3, 1997; accepted February 18, 1998

Purpose. We explore use of "bootstrapping" methods to obtain a measure of reliability of predictions made in part from fits of individual drug level data with a pharmacokinetic (PK) model, and to help clarify parameter identifiability for such models.

Methods. Simulation studies use four sets (A-D) of drug concentration data obtained following a single oral dose. Each set is fit with a two compartment PK model, and the "bootstrap" is employed to examine the potential predictive variation in estimates of parameter sets. This yields an empirical distribution of plausible steady state (SS) drug concentration predictions that can be used to form a confidence interval for a prediction.

Results. A distinct, narrow confidence region in parameter space is identified for subjects A and B. The bootstrapped sets have a relatively large coefficient of variation (CV) (35-90% for A), yet the corresponding SS drug levels are tightly clustered (CVs only 2-9%). The results for C and D are dramatically different. The CVs for both the parameters and predicted drug levels are larger by a factor of 5 and more. The results reveal that the original data for C and D, but not A and B, can be represented by at least two different PK model manifestations, yet only one provides reliable predictions.

Conclusions. The insights gained can facilitate making decisions about parameter identifiability. In particular, the results for C and D have important implications for the degree of implicit overparameterization that may exist in the PK model. In cases where the data support only a single model manifestation, the "bootstrap" method provides information needed to form a confidence interval for a prediction.

KEY WORDS: pharmacokinetic; predict; model; parameter; bootstrap; forecast; algorithm; estimates; simulate; decision support; pharmacometrics.

[1] Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and The Programs Bioengineering and Medical Information Science, University of California, San Francisco, California 94143-0446.

[2] Department of Statistics, Colorado State University, Fort Collins, Colorado 80523-1877.

[3] Department of Clinical Pharmacy, University of California, San Francisco, California 94143-0446.

[4] To whom correspondence should be addressed. (e-mail: hunt@itsa.ucsf.edu)

ABBREVIATIONS: CV, coefficient of variation; PK, pharmacokinetic; SS, steady state; θ, a set of PK model parameter values; C_{it}, an observed experimental drug level for the ith individual; $t = 1, \ldots, T_i$, sampling times; $C_{true,i,t}$, error-free, template (true) drug level data; $j = 1, \ldots, M$, bootstrap sample index; $\bullet^*$, a bootstrapped pseudo-value; $\hat{\bullet}$, parameter estimate; f, a density; $f(C_{it}; \phi_{it})$, a density representing the distribution of plausible drug levels; var(C_{ij}), variance of an observed drug level; Ω, a covariance matrix; z_i, a normal variate; w_{ij}, a weight used in model fitting; N($\bullet$), distributed normally; V, k_a, k_{10}, k_{12}, k_{21}, and t_{lag}, parameters specific to the two compartment PK model; CL, CL_D, V_c, V_T and V_{SS}, physiologic PK parameter designations.

INTRODUCTION

An expected benefit of fitting individual pharmacokinetic (PK) and pharmacodynamic data with models is that one can then make inferences and potentially useful predictions. Of special interest are predictions of "the concentration-time profiles of drugs during chronic drug administration from knowledge obtained in single dose studies" (1). A well known problem arising from such predictions is that the characterization resulting from an acceptable fit of single dose data can, for that individual, yield unrealistic predictions, for example at an hour after dosing during oscillatory steady state (SS) following a fixed interval oral dosing regimen. The absence of some measure of reliability increases risk when such predictions form the basis for decision making. The problem can have any of several root causes. Consider the classical two-compartment model with absorption. Purves (2) identifies the sources of difficulty in parameter estimation which carry through to prediction, including multiple outcomes resulting from the rank order of the rate constants, a variety of local minima, and anomalous outcomes. Purves offers useful strategies to limit such difficulties, but notes that they can not be completely avoided. As a consequence, a fit of a model to an individual data set may be reasonable, yet the fitted parameter values may poorly predict the time trajectory of SS drug levels for that individual. Frustration arising from such experiences has been responsible, in part, for the preference of some investigators for noncompartmental PK parameter summaries, which frequently fail to optimally extract useful information from the data, and severely limit one's ability to predict.

The field will benefit from having techniques to measure the reliability for such predictions. They will not substitute for existing analysis techniques, but may offer means to new insights on the problems under study. They should allow one to state, for example, that at one hour after dosing at SS for the regimen specified, the individual is expected to have a drug plasma level within an estimated range of 36–52 μg/ml, with a probability of 10% that the estimated range fails to cover the true level, assuming system constancy. Such measures can aid decision making. In theory, information contained in the variance-covariance matrix that results from the fitting procedure can be incorporated into a strategy to provide a reliability measure, but such a strategy can be unreliable when problems like those Purves discusses are severe.

Here we present results exploring how bootstrapping methods (3,4) can be used to identify the region in parameter hyperspace that is most consistent with the experimental data *and related knowledge.* The information generated by that process is then used to obtain a measure of reliability of generalized predictions made for new settings of the independent variables. The data-dependent bootstrap propagates uncertainty through the PK model, and is viewed as a simulation-based approach to map data variation to PK parameter uncertainty, and hence to PK model predictions. When additional data is available this propagation can be extended further through a pharmacodynamic model to predictions of response.

There are three main steps.

1. Use a bootstrap to generate single-dose pseudo-observations from the same general region that contains all of the

0724-8741/98/0500-0690$15.00/0

individual experimental data. This step essentially generates equally plausible values to what has been observed.

2. Use each set of real or pseudo-data and established methods to fit PK model parameters. This step essentially maps a plausible region of the PK parameter space in the vicinity of the best fit parameter set.

3. Finally, use these fitted parameter sets to generate a frequency distribution of plausible, generalized predictions for multiple dosing, and use that distribution to measure the reliability of the traditional single value prediction.

Although the methodology is general and extendible to all PK and pharmacodynamic models, we limit study here to data that can be fit to the classical two compartment PK model with first-order absorption, because it is a model that is frequently used and one that poses notorious difficulties.

Consider sets of single dose drug level versus time data, from each of several different subjects. A PK model has been fit to each dataset, and each fit is judged acceptable. Each set of PK parameter values can be used to calculate subject-specific drug levels during a repetitive dosing regimen. Extending the prediction for the new conditions stresses the model. A serious problem, which we illustrate later, is that it is possible that the fitted parameter values for some subjects give reasonable predictions, whereas predictions for other subjects are grossly inaccurate, yet in both cases predictions following a single dose are acceptable. In practice, of course, one never knows true levels. Is there some other means of distinguishing between the former and latter subjects? In our example, visualization of bootstrap results makes the distinction clear, while also providing information on the reliability of an extended prediction.

METHODS

Overview of Methodology

In general, the approach is as follows. For each set of individual experimental drug levels we identify an error model for drug concentration levels, as a function of time, which specifies probability regions like the shaded region in Fig. 1

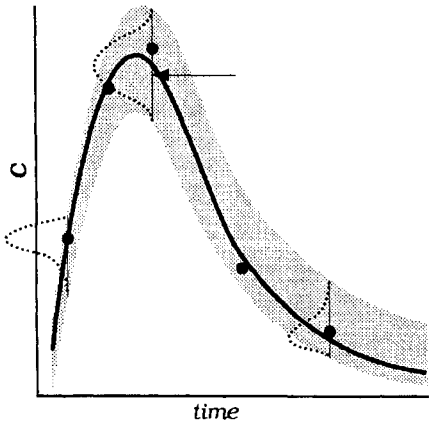

Fig. 1. A schematic of drug level data vs. time. In our simulation example, subjects have 20 observation times. Here only 5 are shown. The solid curve is the drug level predicted from the fit of the PK model to the data. The shaded region represents a model-based high probability region as discussed in the text. The three dashed frequency functions show three different possible parametric bootstrap distributions.

and an appropriate error structure describing how observations are expected to deviate from the true PK model prediction. Figure 1 shows a hypothetical mean concentration curve, and the shaded region might represent, for example, a pointwise 95% probability region. Assume that the PK effects are relatively stable over time, and that one has identified a satisfactory, applicable PK model. The choice of fitting criterion used to produce a curve like that in Fig. 1 clearly depends on the assumed error structure. Often, expert knowledge can support the choice of error structure. In our example, the fitting criterion is to minimize the sum of squared weighted residuals; this follows from a particular error structure we assume below.

A core assumption is that each observed experimental drug level, C_{it}, is a realization from a distribution of plausible values having an estimable variance and a mean specified by the true PK model. Thus the central curve in Fig. 1 shows the time trajectory of the mean concentration, and flanking regions like the one in Fig. 1 are of most interest when they reflect typical variation of individual data. In this case, unobserved values that lie within such a region contain similar information to the observed data, and, in some sense, can be treated as data (5,6). It follows that one may bootstrap the observed data— randomly generate new plausible data based on the observed—to create a pseudo-dataset, and then fit the model to that pseudo-data to obtain a set of parameter pseudo-estimates that provide another estimate of that individual's PK parameter values. The latter is expected to provide predictions that are as plausible as those obtained from the original experimental fit. Such predictions can, upon repetition of the process, reflect, for example, the sampling variability inherent in the original data.

Bootstrapping Pharmacokinetic Data

1. Choose a Model for Bootstrapping the Experimental Data

Three options that are easy to implement are:

(a) Parametric Bootstrap. Assume some density $f(C_{it}; \phi_{it})$ represents the distribution of drug concentrations for individual i at times $t = 1, \ldots, T_i$, where T_i is the total number of observations on individual i. Here, ϕ_{it} is estimated from the data using, for example, maximum likelihood. Denote the estimate as $\hat{\phi}_{it}$. Sample C_{it}^* from $f(C_{it}; \phi_{it})$ to create M pseudo-datasets. This option is expected to be sensitive to the form of f.

(b) Nonparametric Bootstrap. For each i independently, and conditional on each t in turn, let $f(C_{it}; \phi_{it})$ assign mass $1/T_i$ to the concentration equal to the fitted $\hat{C}_{it}$ plus each of the T_i residuals. During bootstrapping, the time points remain unaltered. This approach can be implemented within each individual by resampling the residuals for individual i with replacement, and adding these pseudo-residuals back to the fitted values to create pseudo-data C_{it}^*. Weighting and/or rescaling of the residuals might be needed to obtain a suitable error structure. This method requires weaker assumptions than (a).

(c) Smoothed Bootstrap. Consider each datum to be a pair (t, C_{it}). For each i independently, let $f(C_{it}; \phi_{it}) = f(t, C_i) = 1/T_i; \sum_{j=1}^{T_i} N((t,C_{it})', \Omega)$ where Ω is a covariance matrix that supplies a small amount of noise with suitable error structure. This

approach can be carried out by randomly sampling time-concentration pairs from the existing data, and then, for these pairs, creating each new pair (t^*, C_{it}^*) according to $(t^*, C_{it}^*) = (t, C_{it}) + (z_1, z_2)$ where $(z_1, z_2) \sim N(0, \Omega)$. This option is expected to provide smoother, more intuitive pictures for graphical display.

2. Bootstrap the PK Parameters

Denote by θ the parameters of the chosen PK model, and by $\hat{\theta}$ their estimates based on observed data. For the original experimental data and each pseudo-dataset created above, fit the PK model to obtain the primary parameter estimates $\hat{\theta}$, and the pseudo-estimates $\hat{\theta}_j^*$, respectively, $j = 1, \ldots, M$. Because bootstrapping effectively generates pseudo-data representative of a shaded region like that in Fig. 1, the M sets of $\hat{\theta}_j^*$ are representative of the types of PK response that an individual might have.

3. Simulation

Use $\hat{\theta}$ and the $\hat{\theta}_j^*$ to simulate drug levels specific to the individual. Use graphs, summary statistics and percentiles from bootstrap distributions to draw inferences.

A Specific Example

In this example, we used data for four subjects (A-D) as specified in the next section, and we illustrate the parametric bootstrap option. We set $T_i = 20$ for all i, and $M = 100$. These decisions translate into the following specific steps.

1. The density $f(C_{it}; \phi_{it})$ is intended to represent the shape and dispersion of plausible drug levels that would result through time if repetitive studies could be carried out on the unchanged ith individual. The choice for f used in this study assumes that C_{it} is Gaussian with mean at time t specified by the PK model, and $\text{var}(C_{ij})$ proportional to the mean. Thus, in option (a) we used weighted least squares to fit the model to the data to obtain $\hat{C}_{it}$. Alternative parametric bootstrap models, as suggested at the first and fifth observation in Fig. 1, can be tailored to the individual and/or to the observation time. However, for our data the single model suffices.

The bootstrap proceeds at the first observation time by simulating one hypothetical drug level C_{i1}^* from the bootstrap distribution: $C_{i1}^* = \hat{C}_{i1} + (z_{i1}/\sqrt{w_{i1}})$, where z_{i1} is a normal variate and $w_{i1} = 1/\hat{C}_{i1}$. The w_{ij} are the weights used for model fitting. Repeat the preceding for all additional observation times. The resulting 20 values of C_{it}^* form one of 100 bootstrapped pseudo-datasets of plausible drug levels.

2. Next, while observing Purves' fitting admonitions (2), fit the PK model to each pseudo-dataset, as above, using the Gauss-Newton weighted least squares fitting algorithm option of WinNonlin 1.1 (SCI, Cary, NC). Thus, each set of C_{it}^* results in a fitted $\hat{\theta}_i^*$. By repeating the process M-1 times, we completed the bootstrap methodology for individual A. Individuals B, C and D were treated identically.

3. We use the four sets of 100 $\hat{\theta}_j^*$ as auxiliary inputs to the PK model equations to generate drug level pseudo-predictions. To examine results, we consider making predictions for 1, 3 and 6 hours after dosing at SS for a dosing regimen of 200 mg (see next section) given orally every 12 hours. Therefore, having estimated that the four subjects will reach steady state before the 30th dose, we calculate drug levels at 361, 363 and 366 hours after the first dose. For each individual, the resulting predicted SS levels at any particular time are expected to form a frequency distribution spanning a drug level range that contains the prediction for that time made using the original $\hat{\theta}$. That frequency distribution may be used to form an approximate confidence interval. For example, the range formed after eliminating the five largest and smallest values is an approximate 90% confidence interval.

Experimental Data

The PK data used is based on previously reported data (7) that were collected as part of a classical corporate Phase II clinical trial; drug plasma levels were determined at $i = 20$ times, at 0.083 to 48 hours following a 200 mg oral drug dose. The development team fit the classical two-compartment open model with first order absorption commencing after an individual lag-time to the data. Mean parameter estimates and coefficients of variation (CV, as %) for 22 individuals were V: 4.1 (25%) liters; k_a: 6.0 (77%) hr^{-1}; k_{10}: 0.09 (22%) hr^{-1}; k_{12}: 0.35 (83%) hr^{-1}; k_{21}: 0.36 (64%) hr^{-1}; and t_{lag}: 0.19 (37%) hrs. Based on prior knowledge a bioavailability $\equiv$ 1.0 could be assumed.

As part of an earlier study (7) we simulated a large number of PK parameter sets, retaining the observed covariance structure, where the mean and variance of each parameter were the same as those above. From these, four examples have been selected and designated individual A, B, C and D. We used each set to generate error-free, template (true) drug level data, $C_{true,i,t}$, at the 20 observation times. Error, selected independently at random from a normal density having a mean of zero and a variance chosen to induce a 10% CV for C_{it}, was then added to each $C_{true,i,t}$ to obtain a simulated experimental drug level, C_{it}. Using the above WinNonlin fitting algorithm and following the producer's instructions (8), we fit the PK model to each of the four data sets using the above weighting scheme. We added the constraint that all parameter estimates be positive and within the following ranges: V: 1-25 liters; k_a: 0.1-50 hr^{-1}; k_{10}: 0.01-1.0 hr^{-1}; k_{12}: 0.001-100 hr^{-1}; k_{21}: 0.001-100 hr^{-1}; and t_{lag}: $\leq$ 0.5 hrs.

Each fit gave a plausible concentration time curve. The bootstrap methodology was then applied to each of the four sets of C_{it}. A discussion of fitting methodology and quality of fit are provided in (9) for three similar data sets.

Physiologic Parameters

Many investigators prefer to parameterize PK models using physiologic parameters. For each bootstrapped set of pseudo-data we also calculated an alternative set of parameter pseudo-estimates: CL^*, CL_D^*, V_C^*, V_T^* and V_{ss}^*; where the notation is as defined by Jusko (10). For the majority of the discussion to follow, however, we use the micro constants listed in the previous section because that was the parameterization selected by those who generated the original data. Doing so also facilitates comparison between this report and several cited references.

RESULTS

Figures 2–5 show the primary parameter estimates, the 100 bootstrap estimates, and the true values, in sets of pairwise

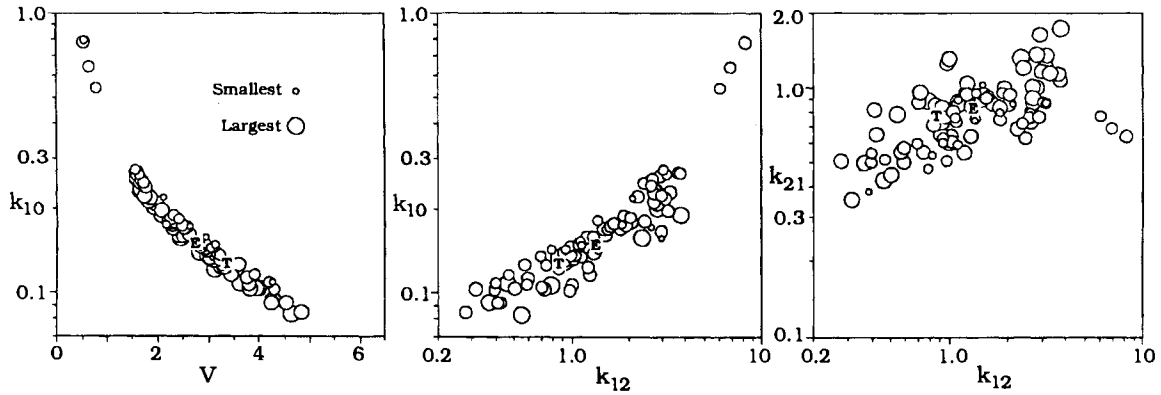

Fig. 2. Paired scatter plots of four PK parameters for subject A. One circle is shown for each point. For each subject the area of the circle corresponds to the predicted SS drug level. From the smallest to the largest, circle area increases through 20 equal-increment steps. In some cases an axis has a log scale. The **E** designates the fitted, experimental primary parameter estimates. The **T** designates the corresponding true values. In the left and center plots one of the 100 bootstrap sets is not shown (is outside the plot frame): (k_{10}, V) = (1.72, 0.20) and (k_{10}, k_{12}) = (1.72, 19.7).

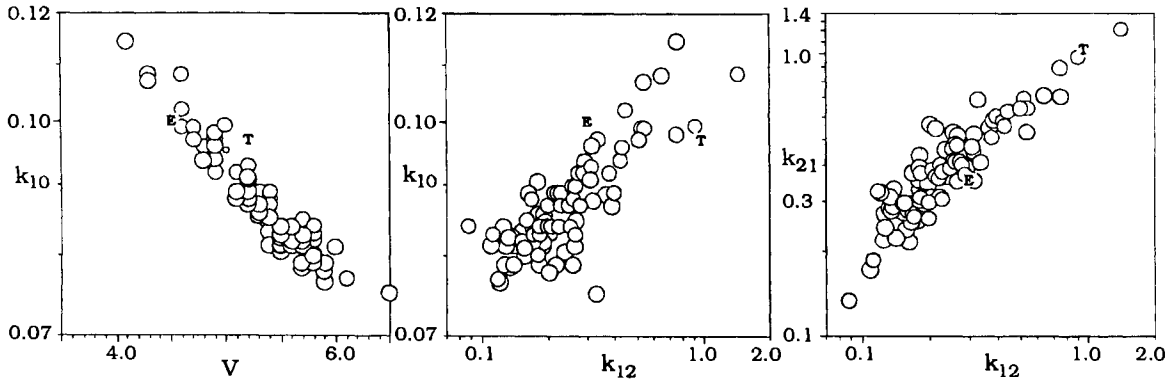

Fig. 3. Paired scatter plots for subject B as described in Fig. 2. If the minimum value in Fig. 6B is deleted, then the range of circle sizes looks more like that in Fig. 2. Six of 100 sets are not shown in the left and center plots; one is not shown in the right plot. They are: (k_{10}, V) = (0.052, 11.6), (0.22, 2.7), (0.38, 1.2), (0.70, 0.61), (0.72, 0.61), (1.78, 0.30); (k_{10}, k_{12}) = (0.052, 0.008), (0.22, 46.0), (0.38, 5.4), (0.70, 7.9), (0.72, 5.7), (1.78, 22.8) and (k_{21}, k_{12}) = (15.7, 46.0).

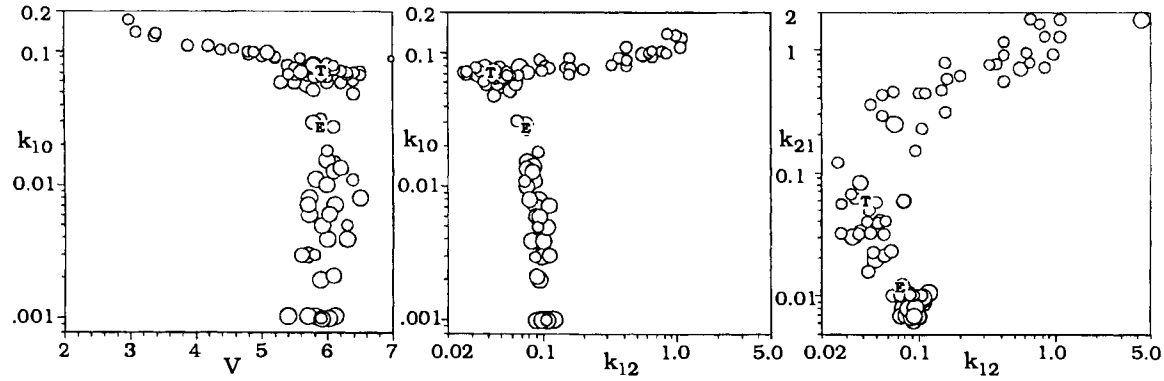

Fig. 4. Paired scatter plots for subject C as described in Fig. 2. One of the 100 sets is not shown in the left plot; two are not shown in the right plot. They are: (k_{10}, V) = (0.063, 25.0) and (k_{21}, k_{12}) = (93.5, 52.2), (4.9, 7.64).

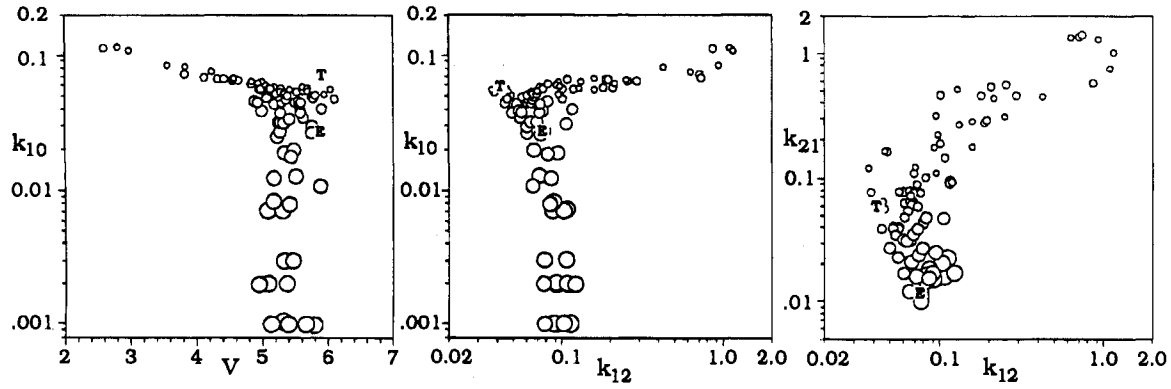

Fig. 5. Paired scatter plots for subject D as described in Fig. 2. All 100 parameter sets are represented in each plot.

scatterplots for each individual; the fits are adequate. In these plots, one set of parameter values is indicated by the location of each circle. The area of the circle corresponds to predicted SS drug level as discussed below. The agreement or lack of agreement between truth and estimates in these plots is consistent with observations from the simulation study of Laskarzewski, *et al.* (11). Note that to better visualize the data the scales for k_{10}, k_{12} and k_{21} are logarithmic in all four plots, whereas those for V are linear. Summary statistics for each are listed in Tables I and II. To emphasize the differences in steady state predictions, the area for each circle in Figs. 2–5 corresponds to predicted SS drug level, with the largest area corresponding to the maximum of the 100 simulated values *for each individual*, and the smallest area corresponding to the minimum.

Figure 6 shows the estimated, bootstrap, and true SS drug levels for each individual. The predicted SS drug levels were within 3% of the template levels for subjects A and B. However, the corresponding levels for subjects C and D were off by as much as 90%. There was no clear indication from either the experimental data or the results of the fit that such discrepancies

Table I. Summary Statistics for 100 Bootstrap Estimates for Subjects A and B

		PK parameter pseudo-estimates					Drug levels time after dosing at SS		
		V	k_a	k_{10}	k_{12}	k_{21}	1 hr	3 hr	6 hr
Mean	A	2.7	9.8	0.18	1.76	0.81	65.6	46.4	39.0
	B	5.2	8.9	0.10	0.46	0.42	53.8	42.2	34.4
Median	A	2.8	7.5	0.15	1.31	0.78	65.3	46.7	39.3
	B	5.4	7.6	0.09	0.22	0.38	54.8	42.5	34.4
CV(%)	A	35	90	64	77	31	5	5	6
	B	24	176	92	239	46	9	6	3
Minimum	A	0.5	1.4	0.08	0.28	0.35	58.2	41.9	34.0
	B	0.3	0.9	0.05	0.01	0.07	21.3	20.8	31.4
Maximum	A	4.8	49.2	0.79	8.17	1.71	74.8	50.8	43.2
	B	11.6	49.3	0.72	7.91	1.18	58.2	45.0	36.6

Table II. Summary Statistics for 100 Bootstrap Estimates for Subjects C and D

		PK parameter pseudo-estimates					Drug levels time after dosing at SS		
		V	k_a	k_{10}	k_{12}	k_{21}	1 hr	3 hr	6 hr
Mean	C	5.6	5.5	0.06	0.083	1.31	74.2	66.6	57.7
	D	5.1	7.1	0.04	0.16	0.18	105.8	96.7	86.7
Median	C	5.8	4.4	0.07	0.09	0.04	60.7	52.4	42.8
	D	5.3	5.3	0.05	0.83	0.06	84.1	75.6	65.5
CV(%)	C	15	912	80	636	736	34	38	43
	D	13	89	60	141	162	39	44	49
Minimum	C	1.8	1.5	0.01	0.03	0.07	22.4	20.7	33.2
	D	2.6	1.6	0.00	0.04	0.01	58.3	58.4	49.2
Maximum	C	7.0	40.3	0.25	52.20	93.50	118.7	111.8	103.1
	D	6.1	37.9	0.11	1.15	1.34	213.9	205.9	196.6

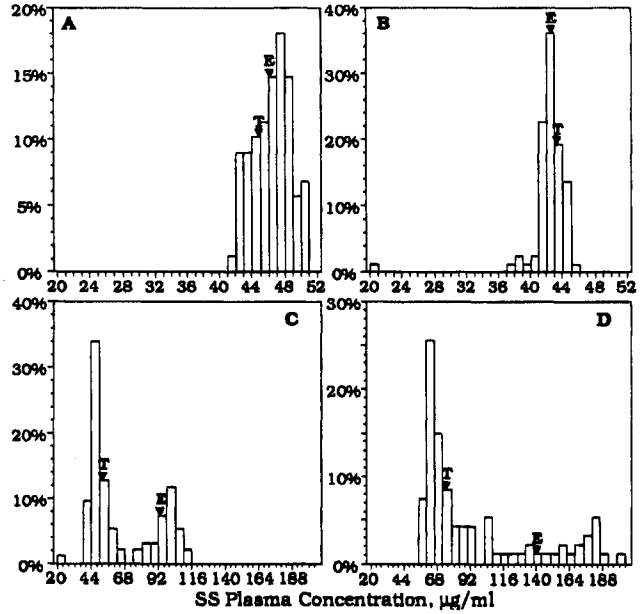

Fig. 6. Frequency histograms for simulated predicted drug levels 3 hours after dosing at steady-state. **E** indicates the value estimated using the fitted (experimental) primary parameter estimates, and **T** indicates the corresponding true value.

might be anticipated. The distribution of bootstrap estimates, however, is clearly qualitatively different for subjects A/B than for C/D.

In addition to the differences shown in Figs. 2–5, inspection of all possible bivariate scatter plots for the corresponding physiologic parameter sets (not shown) also revealed two sets of distinctly different patterns, one for subjects A and B, and the other for subjects C and D. Clear evidence of qualitative differences, namely patterns similar to those seen in Figs. 4 and 5, were evident in the bivariate scatter plots of subjects C and D, specifically (CL_D^*, CL^*), (V_T^*, V_C^*) and (CL_D^*, V_{SS}^*).

Subjects A and B

The 100 bootstrap parameter estimates for subject A map a well-defined five dimensional (5D) density (excluding t_{lag}), at least as viewed in bivariate scatter plots like those in Fig. 2. The marginal distribution for each rate constant, however, has a large range and a relatively large CV (Table I). The functional dependencies that are inherent in the PK model equations are reflected in the observed correlations. Each of the conditional distributions for the three pairs shown—($k_{10}|V$), ($k_{10}|k_{12}$) and ($k_{21}|k_{12}$)—have relatively small conditional ranges and conditional standard deviations. Results for subject B are similar to those for A.

Given the large CVs of each parameter, it is striking that the SS drug levels for A in Fig. 6 and Table I are tightly clustered, with CVs of only 5–6%. This and the similar results for B might seem counter-intuitive. The explanation is the fact that each value within a parameter set is strongly conditioned on each of the other parameter values in that set. Table 3, showing a cross-correlation matrix of the PK parameter pseudo-estimates, confirms this dependency. Thus, a modest random change in one or more experimental drug level may give rise to a second set of fitted parameter values that can seem quite different from the first when only values of a specific parameter are compared, whereas SS predictions may be similar. Such differences can be disconcerting.

Subjects C and D

Although the sets of bootstrap PK parameter estimates for subjects A and B produced a homogeneous set of SS drug level predictions, the results for C and D were dramatically different. That difference is emphasized by the clustering of larger and smaller circles in Figs. 4 and 5. Clearly, there is evidence in Fig. 6 of bimodality in the SS simulations for C, and possibly D. Unlike for A and B, different SS levels for C and D tend to correspond qualitatively to different regions of the parameter space. Note that the circles appear to fall in two clusters: one of mostly smaller circles and a second of mostly the larger ones. This shows that the data support two conflicting model manifestations and corresponding future predictions. In this context a "model manifestation" is a distinct region of predicted response that may correspond to a specific region of parameter space. However, if small variations in parameter space tend to produce predictions that vary between quite different regions of prediction response space (as occurs for subjects C and D in our example), then the word "manifestations" is appropriate for indicating that model predictions effectively arise from distinct competing estimates of the modeled process. This uncertainty also reduces the correlation in Table 3. Also note that the fitted, experimental parameter values for both C and D fall in a region away from the truth, and away from the regions in Figs. 2 and 3.

Measuring Reliability

The parameter summary statistics across the bootstrap sets (Table I and II) provide a useful measure of parameter uncertainty, but in practice it is the predicted drug levels that are of primary interest. In Guzy and Hunt (9), similar data, along with quality of fit data is compared to information from the variance-covariance matrix resulting from the fit of three similar data sets from the same PK protocol.

Together, the bootstrapped parameter sets and SS drug levels may contain sufficient information to provide a measure of reliability for a specific prediction made using $\hat{\theta}$ (4). For example, by eliminating the 5 largest and smallest values from the predictions for A we obtain these estimates for a 90% confidence region at 1, 3 and 6 hours after dosing, respectively: 59.9–71.6, 42.4–50.2, and 35.4–43.0 µg/ml. Although the bootstrap might be used to estimate confidence regions for PK parameters or predicted SS levels by taking appropriate quintiles of the pseudo-estimates, such an effort would not account for the larger uncertainty arising from the discovery that two qualitatively different estimates of reality are supported by the analyses for C, and possibly D.

All Four Subjects

Why are the predictions for C and D so strikingly different from those of A and B, even though the original fits of the experimental data were all judged acceptable? The band of plausible drug levels (as in Fig. 1) for C and D can apparently

Table III. Cross-Correlation Matrix of Bootstrap Estimates for Subjects A and C

Subject A	V	k_a	log k_{10}	log k_{12}	log k_{21}
V	1				
k_a	0.62	1			
log k_{10}	−0.95	−0.54	1		
log k_{12}	−0.95	−0.61	0.91	1	
log k_{21}	−0.43	−0.40	0.31	0.65	1
Subject C	**V**	**k_a**	**log k_{10}**	**log k_{12}**	**log k_{21}**
V	1				
k_a	0.15	1			
log k_{10}	−0.37	0.07	1		
log k_{12}	−0.61	0.15	0.24	1	
log k_{21}	−0.46	0.13	0.74	0.71	1

accommodate at least two different manifestations of the two compartment model, a property that warrants further research.

In Fig. 7, we plot the pseudo-estimates, pooled across all four subjects; SS estimates greater than 72 μg/ml for C and 108 μg/ml for D are shown with dark circles. The within-subject conditional relationships are observed to extend across all four subjects. The pooled data represented by the open circles appear to map a large, well-defined 5D relationship, one that we suggest represents one manifestation of the two compartment model. The parameter sets corresponding to dark circles are believed to represent a portion of a different manifestation, one that branches from the one identified above.

DISCUSSION

The results are both interesting and apparently informative, suggesting that the bootstrap can shed light on the issues of parameter identifiability, as well as providing approximate confidence intervals for predictions. Of the results generated, the most striking are the differences between subjects A/B and C/D, especially since the fits to all four experimental data sets appeared reasonable. Taken together, the results indicate that a prediction made using the best fit parameter estimates for subjects C and D should be treated as being less certain than corresponding predictions for subjects A and B. The bootstrap is thus a useful aid for PK decision making.

As demonstrated, bootstrapping can identify when two (or more) model manifestations are plausible characterizations of the data, but it alone may not identify one as being the more preferred representation. Having additional information, however, may allow one to favor one manifestation over another. Suppose, for example, that only one of a dozen individual data sets shows evidence of two or more PK model manifestations, and that application of the methodology to that data set has given results similar to the data for C. One might be able to draw strength from the other eleven individuals, perhaps by revising the parametric bootstrap distribution in a Bayesian fashion, to focus inference on the more plausible manifestation.

Although each of the PK parameter pseudo-estimates is solved using the same model, fitting procedure, and starting value, the data differ in each case. Suppose that, for data in a banded region like that in Fig. 1, the likelihood has a form with two competing local maxima, and that the potential variation in the data represented by such a band is sufficient to vary the heights of these maxima, but not to change fundamentally the shape of the likelihood. If this is the case, it could cause the apparent presence of two model manifestations, as is seen for

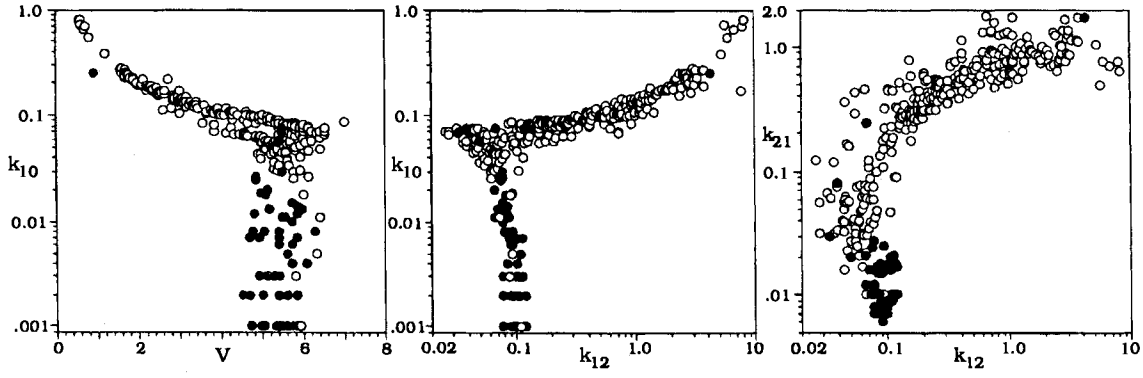

Fig. 7. The data in Figs. 2–5 are pooled and replotted. Points roughly corresponding to a second PK model manifestation (see explanation in text) are shown with dark circles. The main body of data for subjects A-D (open circles) gives the appearance of belonging to a larger, more general multivariate density that may represent a single, primary, drug specific manifestation of the PK model.

subject C. However, the results alone are not conclusive proof that multiple potential local maxima exist.

To our knowledge Laskarzewski, *et al.* (11), and later Betzien, *et al.* (12), were among the first to report evidence of two or more PK model manifestations, whereas much has been published on the "flip-flop" phenomenon. They show that different sets of drug levels simulated from the same error-free template data from a two compartment PK model with absorption, can give rise to a bimodal set of fitted parameter values. Laskarzewski, *et al.* observed such bimodality when the error added was 10%, but not when it was 5%. It makes sense that with increasingly precise data the problem of structural identifiability (13–15) of the parameter values decreases. With very imprecise data, the greater uncertainty can support a variety of manifestations of the same model, and hence the problem of structural identifiability increases.

What about different models? Is it possible that for some sets of bootstrap data one may get a better fit using a different model from the same general class, in this case a one or a three compartment model with absorption? Based on the research of Godfrey and Chapman (14) and Woodruff, *et al.* (15,16), the likely answer is yes. Given that, one should be able to use simulation approaches to provide new, quantitative information about relative model indistinguishability for a given experimental data set.

The performance of the bootstrap is highly dependent on the nature of the available data, and on the assumptions that are imposed. This is particularly true for the parametric bootstrap approach; the other bootstraps listed are less model-dependent. Obtaining a sample of simulated predictions with only a 5% coefficient of variation (subject A, Table I) can not therefore be taken as a signal that the simulated predictions are accurate, or that they accurately reflect reality, since the PK model may be somewhat misspecified.

When bootstrapped SS levels are within the range typical of experimental error and hence might be treated as being experimentally indistinguishable, should the corresponding parameter sets also be treated as being indistinguishable? When inspecting parameter sets one tends to focus on differences in one and two dimensions. In so doing one tends to forget that the parameters are part of a single vector in parameter hyperspace. Careful plotting of bootstrap pseudo-estimates allows one to identify a region in parameter space that can give rise to similar, even experimentally indistinguishable predictions. Although the pseudo-estimates permit visually appealing plots that may aid decision making, they contain no more information, in a statistical sense, than the original data. However, estimating individual PK parameter values from experimental data is only one step in the analysis process, and our results show that (visual) identification of the region in parameter hyperspace that is most consistent with the entire collage of experimental data is equally important and can lead to new insights, such as those obtained here.

ACKNOWLEDGMENTS

The authors thank Lewis Scheiner, Dan Weiner, Robert Upton, Ronald Siegel, David Bourne, Svein Øie, and two anonymous reviewers for their very helpful comments, useful insights and suggestions. The work was supported in part by grant GM-08388 from NIGMS and a UCSF Instructional Technology Award.

REFERENCES

1. G. Levy. Applied pharmacokinetics—a prospectus, in W. E. Evans, J. J. Schentag, and W. J. Jusko (eds.) *Applied Pharmacokinetics, Principles of Therapeutic Drug Monitoring,* Third Ed., Applied Therapeutics, Vancouver, WA, pp. P.1–P.4 (1992).
2. R. D. Purves. Multiple solutions, illegal parameter values, local minima of the sum of squares, and anomalous parameter estimates in least-squares fitting of the two-compartment pharmacokinetic model with absorption. *J. Pharmacok. Biopharm.* **24**:79–101 (1996).
3. B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap, Monograph on Statistics and Applied Probability, No. 57.* Chapman and Hall, New York, pp. 153–190 (1993).
4. M. T. Markus. *Bootstrap Confidence Regions in Nonlinear Multivariate Analysis,* DSOW Press, Liden, pp. 41–65 (1994).
5. A. P. Dempster. Probability, evidence, and judgment, in J. M. Bernardo, *et al.* (eds.) *Bayesian Statistics 2,* Elsevier Science Publishers, Amsterdam, pp. 119–131 (1985).
6. S. French. Group consensus probability distributions: a critical survey, *ibid.,* pp. 157–182.
7. S. Guzy and C. A. Hunt. Validation of a decision support system for use in drug development: Pharmacokinetic data. *Pharm. Res.* **14**:1287–1297, 1997.
8. J. Gabrielsson and D. L. Weiner. *Pharmacokinetic And Pharmacodynamic Data Analysis: Concepts and Applications,* Swedish Pharmaceutical Press, Upsula (1994).
9. S. Guzy and C. A. Hunt. Measures of uncertainty for pharmacokinetic and pharmacodynamic parameter estimates: A new computerized algorithm. *Comp. Biomed. Resh,* **29**:466–481 (1996).
10. W. J. Jusko. Guidelines for collection and analysis of pharmacokinetic data, in W. E. Evans, J. J. Schentag, and W. J. Jusko (eds.) *Applied Pharmacokinetics, Principles of Therapeutic Drug Monitoring,* Third Ed., Applied Therapeutics, Vancouver (WA) pp. 2.9–2.29 (1992).
11. P. M. Laskarzewski, D. L. Weiner and L. Ott. A simulation study of parameter estimation in the one and two compartment models. *J. Pharmacok. Biopharm.* **10**:317–334 (1982).
12. G. Betzien, B. Kaufman, B. Schneider and W. A. Ritschel. Simulation studies of error of parameter estimates in pharmacokinetics. *Arzneim-Forsch./Drug Resh.* **35**:7–14 (1985).
13. S. Vajda, K. R. Godfrey and M. J. Chapman. Similarity transform approach to identifiability analysis of nonlinear compartment models. *Math. Biosci.* **93**:217–248 (1989).
14. K. R. Godfrey and M. J. Chapman. The problem of model indistinguishability in pharmacokinetics. *J. Pharmacok. Biopharm.* **17**:229–267 (1989).
15. T. J. Woodruff, F. Y. Bois, D. Auslander and R. C. Spear. Structure and parameterization of pharmacokinetic models: Their impact on model predictions. *Risk Anal.* **12**:189–201 (1992).
16. T. J. Woodruff and F. Y. Bois. Optimization issues in physiological toxicokinetic modeling. *Toxicol. Let.* **69**:181–196 (1993).