Validation of a Decision Support System for Use in Drug Development: Pharmacokinetic Data

Serge Guzy[1] and C. Anthony Hunt[2,3]

Received July 3, 1997; accepted July 16, 1997

Purpose. Single dose pharmacokinetic data from several individuals can be used to predict the fraction of the population that is expected to be within a therapeutic range. Without having some measure of its reliability, however, that prediction is only likely to marginally influence critical drug development decision making. The system (Forecaster) described generates an approximate prediction interval that contains the original prediction and where, for example, the probability is approximately 85% that a similar prediction from a new set of data will also be within the range. The goal is to validate that the system functions as designed.

Methods. The strategy requires having a *Surrogate Population* (SP), which is a large number (≥1500) of hypothetical individuals each represented by set of model parameter values having unique attributes. The SP is generated so that a sample taken from it will give data that is statistically indistinguishable from the available experimental data. The automated method for building the SP is described.

Results. Validation studies using 300 independent samples document that for this example the SP can be used to make useful predictions, and that the approximate prediction interval functions as designed.

Conclusions. For the boundary conditions and assumptions specified, the Forecaster can make valid predictions of pharmacokinetic-based population targets that without a SP would not be possible. Finally, the approximate prediction interval does provide a useful measure of prediction reliability.

KEY WORDS: pharmacometrics; pharmacokinetics; simulate; predict; validate; clinical trial; population; decision support; informatics; bootstrap; clinical outcomes; algorithm.

INTRODUCTION

There is a need for expert decision support systems that are designed to increase the efficiency and productivity of the clinical phase of drug development. They will allow investigators to identify successful and unsuccessful drug candidates earlier in the process by making optimum use of accumulated pharmacokinetic (PK) and pharmacodynamic (PD) knowledge. They will deterministically link knowledge from the molecular to the population level. The core of such systems will need to be a knowledge-based representation of a real population. Toward that end, assume that a core database is structured to represent a *Surrogate Population* of hypothetical subjects, where each is characterized by unique attributes and is capable of providing PK and PD data of the same type that is collected during clinical trials. To illustrate its function, assume that some PK, PD and covariate data are available from a random sample of n individuals (taken from the "real" population). For the same independent variable settings, the same type of data is obtained from m ($m \geq n$) randomly selected surrogate subjects. The two data sets are then scrutinized for significant distinguishing characteristics. None are found, either in terms of mean results, the variability within and between subjects, the time course for relevant observations (drug plasma levels, measures of response, etc.), or the relationship between the experimental results, prior knowledge and relevant covariates. The two data sets are statistically indistinguishable. It follows that the Surrogate Population (SP) may, within limits, prove to be a valuable decision support resource to predict how the *real* population is expected to respond. For example, subjects from the SP can be used to simulate the results of candidate clinical trial designs with the aim of selecting the most optimum. One could use the results to evaluate projected therapeutic *population targets* of the type: The fraction F of the population that should exhibit a significant therapeutic effect for a specified dose, formulation, regimen, etc. is $F \geq 0.85$, whereas the fraction having an undesired side effect is $F' \leq 0.05$.

We recently described the design and demonstrated the use of a prototype system (1), the Forecaster, that defines and generates SPs. The Forecaster uses individualized PK/PD model parameter estimates plus covariate data as input variates for a stochastic modeling protocol. It generates unique, smooth, empirical multivariate densities that are intended to reasonably simulate observed interindividual variability, but should not be confused with the classical population based PK/PD approaches (2,3). The Forecaster is designed to extract as much information as possible out of the available data, and to evolve to deliver additional specific capabilities. One key capability is that predictions are accompanied by a measure of their reliability. The objective of this study is to validate this reliability measure for predictions of PK-based population targets. Having a prediction reliability measure allows one to access the benefits of decision analysis (4).

Clearly, the value and utility of such a system can only be judged following several rounds of rigorous scientific testing and validation (5,6). A pivotal part of this process will be to use existing early clinical trial data to predict later existing data for a specific drug. First, however, it is necessary to confirm that the prototype functions at the PK level as designed. Rather than the typical population mean result, we aim to predict the

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

[2] Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, Division of Biomedical Informatics and The Program in Bioengineering, The University of California, San Francisco, California 94143-0446.

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

ABBREVIATIONS: API(s), Approximate prediction interval(s); CV, Coefficients of variation; *MI* and *AMI*, Matching index and approximate matching index; OSS, Oscillatory steady state; PK, pharmacokinetic; PD, pharmacodynamic; SP(s), Surrogate Population(s); REP, Reference Experimental Population; V, k_{01}, k_{10}, k_{12}, k_{21} and t_{lag}, Parameters in the PK model; F, Bioavailability; x, Independent variables; θ, a p-dimensional vector of variates and covariates; β_1*, β_1, β_2, . . . , β_p, A variate or covariate; β_1*, β_1, β_2, . . . , β_p, A specific value of a variate or covariate; $p(\theta)$, Any probability density function; $p(\theta)_S$, The probability density function for a Surrogate Population based on a specific test sample; C_R, Error-free drug level in the "Reference" data set; C_E, Simulated experimental drug level; ϵ, Random error with an expected value of zero.

0724-8741/97/1000-1287$12.50/0

fraction of a population that is expected to have drug level values within a therapeutic range at steady state, which is our *population target*. We use known data to generate the SP. It is then used to predict each population target value. To gain a measure of reliability of that prediction we bracket it with a *approximate prediction interval* (API). The API is a "confidence interval" such that, under ideal conditions, the probability is approximately 85% (say) that the actual population target value is contained therein. We then repeat the entire process 300 times, and by demonstrating that the API actually captures the population target value approximately 85% of the time we show that the Forecaster can make valid PK-based predictions.

METHODS

The Validation Strategy

This is the first in a series of exercises that will focus on the SP approach (Fig. 1) to forecasting. Here we consider only PK data. In subsequent reports we will expand the exercise to PK and PD data, and later will include covariate data. We follow the two part validation strategy outlined in Fig. 2, which is based in part on the decision support system testing and validation plan developed by Sailors, *et al.* (5). As input the Forecaster needs a set of fitted PK parameter values. One also needs the prediction question(s) and specified independent variable conditions (Fig. 2). Harrell, *et al.* (6) discuss that the predictive accuracy of systems should be validated using a bootstrap resampling protocol. Such a protocol is a key component of both Forecaster (Fig. 1B) and of Part II of the strategy. Consequently, the output of many different decision pathways are repetitively tested during each part of the validation process.

The Reference Standard

A large set of data is needed to serve as a *reference standard*. In this case PK parameter estimates from several hundred individuals would be needed. Because no such data set is available, one must be generated, but it should be sufficiently realistic to challenge both the Forecaster and the validation strategy. To build the reference standard we used actual experimental PK data as a template. That data (1) was collected as part of a classical corporate Phase II clinical trial; the development team fit the data to a classical two-compartment, open model with first order absorption commencing after an individual lag-time. Mean parameter estimates and coefficients of variation (CV, as %) for 22 individuals were V: 4.1 (25%) liters; k_{01}: 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 of $F \cong 1.0$ could be assumed.

The details for generating our reference standard are given in the Appendix. Briefly, to add realistic structured interindividual variability we imposed correlation between three parameter pairs, k_{10} vs V, k_{12} vs k_{10} and k_{21} vs k_{12}. The other three parameters were specified to be independent and the frequency distribution for each was selected at random from Fig. 3. We then generated 1500 "reference" PK parameter sets (Fig. 2A) such that their summary statistics were approximately the same as those above. We used each set to generate single dose drug level data and then added random error to each value. The result is 1500 sets of *reference experimental* drug level data. We fit each set to the same PK model used for the template data. These 1500 fitted parameter sets represent an infinitely large *Reference Experimental Population* (REP) and they serve as our reference standard in Part II.

Fitted PK parameter estimates obtained from some drug level data can be unreliable, and some predicted dependent variable output can be invalid (7–10). Based on the simulations of Laskarzewski, *et al.* (11) and the recent research of Purves (9), there is a risk that some parameter sets within the REP will contain invalid parameter estimates. In part to help assess the degree to which this may occur, we selected population targets during oscillatory steady state (OSS), rather than following a single oral dose, the condition for the reference experimental drug levels.

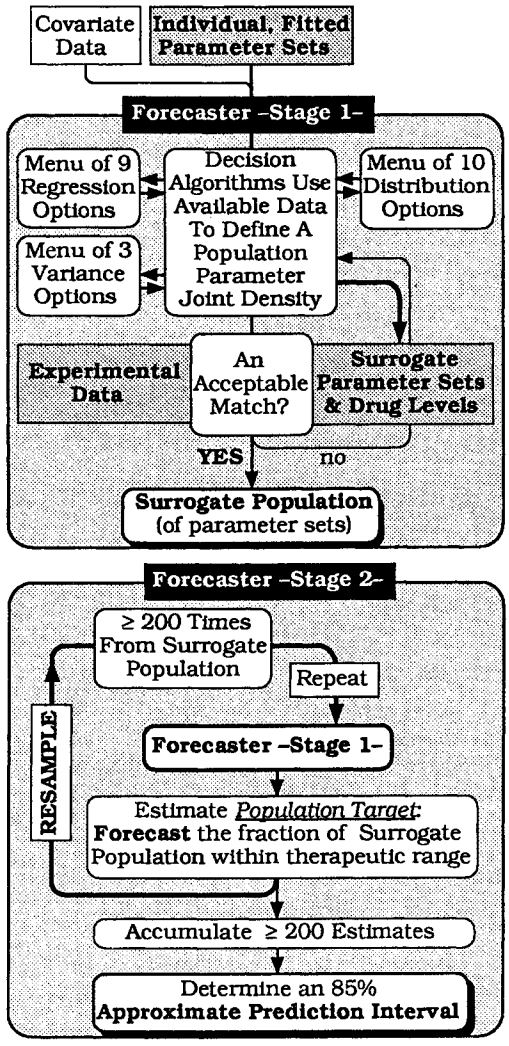

Fig. 1. Forecaster Function. In Stage 1 the Forecaster takes fitted model parameter estimates and follows a protocol (Appendix) to generate and accept a Surrogate Population. In Stage 2 the Forecaster uses the SP to make predictions and calculate an approximate prediction interval (API) for that prediction.

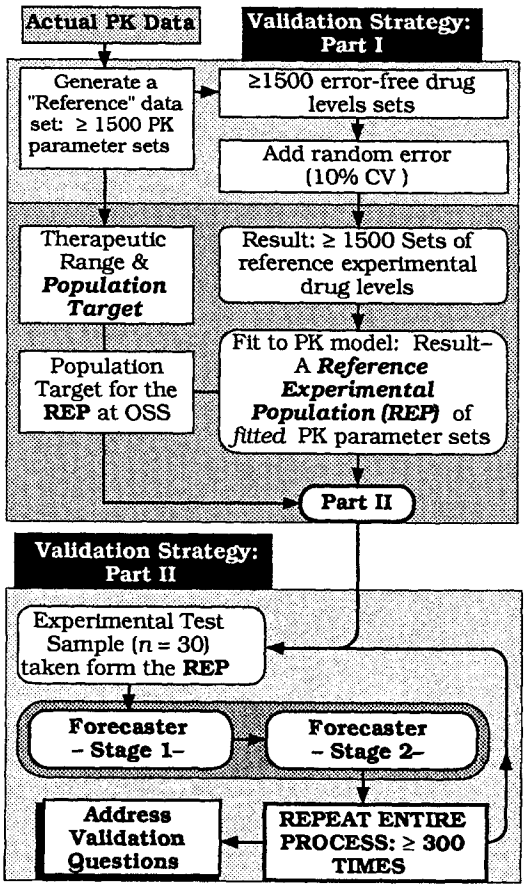

Fig. 2. Validation Strategy. Part I: A schematic representation of the strategy used to generate the Reference Experimental Population, which serves as the reference standard for the validation. In Part II an experimental test sample of $n = 30$ parameter sets is taken at random from the REP and processed by the Forecaster (Fig. 1); that process is then repeated 300 times. OSS: oscillatory steady state.

The Forecaster and the Surrogate Population

The first step in Part II (Fig. 2B) is to obtain a *test sample*. If it is too small, e.g., 5, it will contain essentially no information on interindividual variability. If it is large, e.g., 200, the reliability of a prediction may be acceptable, whereas the probability of actually getting such a large sample may be so small that there is little practical interest in a resulting validation. Here we use the 30-member *test sample* described in Table II. Hereafter that will be the size of all test samples unless specified otherwise.

During Stage 1 the Forecaster automatically evaluates the *test sample* and completes a unique characterization of the PK parameter space that does not require an assumption of normality or log-normality for parameter distributions. It then creates a table of at least 1500 random sets taken from the SP. All subsequent reference to a SP is referring these 1500 sets. A strength of this approach is that even if a given parameter set in the test sample is a poor predictor for that individual, it is likely be quite adequate for another individual in the population. Therefore, absent other information, all test sample sets are weighted equally. This concludes Part I.

The prototype design stipulates that all SPs that meet the criteria (below) will be multivariate distributions where each

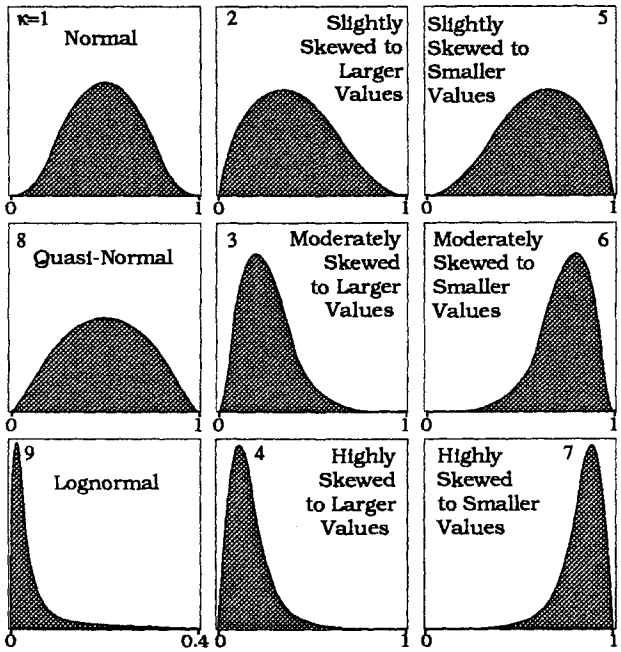

Fig. 3. Menu Options For Density Functions. To facilitate comparisons, the density function options listed in Table 1 are shown. Each option, with the exception of the Lognormal, is generated from a beta function, and in its menu form has limits of 0 and 1. The areas shown are the same.

of p random variates is a PK/PD model parameter (or a covariate). The parameter values for one individual form one p-dimensional vector. We note that all distributions of interest can be generated using the conditional distribution approach (12). The beauty of this approach is that it reduces the task of generating a hypothetical p-dimensional distribution into a series of easily managed univariate generation tasks. We utilize the menus in Fig. 1 to manage univariate generation. Presume that $\theta = (\beta_1, \beta_2, \ldots, \beta_p)$ represents a p-dimensional vector of interest, where β_1, for example, represents all values of the PK volume of distribution. The conditional approach involves the following steps.

1. Generate a value $\beta_1 = \beta_1$ from the marginal distribution of β_1, $\pi(\beta_1)$.

Table I. Regression Function Menus Used to Characterize *Surrogate Populations*

Menu item κ	Regression functions[a]
1	$h_1(k_{1*}) = ak_{1*} + b$
2	$h_2(k_{1*}) = a(k_{1*})^2 + bk_{1*} + c$
3	$h_3(k_{1*}) = a + b(k_{1*})^{-1}$
4	$h_4(k_{1*}) = (ak_{1*} + b)^{-1}$
5	$\log h_5(k_{1*}) = ak_{1*} + b$
6	$h_6(k_{1*}) = a\sqrt{k_{1*}} + b$
7	$h_7(k_{1*}) = a(\log_{10}k_{1*}) + b$
8	$h_8(k_{1*}) = a(k_{1*})^b$
10	$h_{10}(k_{1*})$ = User Defined

[a] One is used to characterize each conditional distribution (see text). All parameter values are positive, and $E[k_k|k_{1*}] = h_\kappa(k_{1*})$, where k_{1*} is the parameter that is correlated with k_k, $k = 1$ to 6.

Table II. Descriptive Statistics for the Experimental Test Sample

	PK parameter					
	V, liters	k_{01}, hrs^{-1}	k_{10}, hrs^{-1}	k_{12}, hrs^{-1}	k_{21}, hrs^{-1}	t_{lag}, hrs
Mean[a]	4.1	6.3	0.097	0.50	0.51	0.19
Std. Dev.	0.989	2.74	0.030	0.454	0.456	0.058
Variance	0.977	7.49	0.009	0.206	0.208	0.0034
CV (%)	24.1	43.4	30.8	90.9	89.0	30.6
Skewness	−0.70	0.38	1.04	1.02	0.92	0.47
Percentile						
10	2.9	2.4	0.066	0.06	0.07	0.12
50	4.2	6.1	0.093	0.37	0.37	0.20
90	5.3	9.9	0.134	1.15	1.13	0.26

[a] For 30 PK parameter sets taken at random from the REP (Reference Experimental Population) and used by the Forecaster to generate the SP.

2. Generate a value $\beta_2 = \beta_2$ from the conditional distribution of β_2 given a value of β_1, $\pi(\beta_2|\beta_1)$.

3. Generate a value $\beta_3 = \beta_3$ from the conditional distribution of β_3 given a value of β_2 and a value of β_1; and so forth through p steps.

We note that most PK/PD data of interest will have rather large interindividual variances. Also, the distribution characteristics of a majority of PK/PD data will represent a small subset of all distribution characteristics. Therefore, we argue that the above tasks can be easily managed using menus having a limited number of choices. Upon completion of the automated process one has the unique characterization of the PK parameter space that is needed. For details see the Appendix.

Three criteria are satisfied before a SP is accepted. First, random samples of ≥30 PK parameter sets from the SP are statistically indistinguishable from the test sample for all model PK parameters. Second, expected drug levels from the SP will be similarly statistically indistinguishable from the reference experimental drug levels. Finally, for a specific sampling time, the range of drug levels predicted from the SP will include, preferably, all of the reference experimental drug levels, and will have a mean and variance similar to these values. All such testing can be done automatically as part of Forecaster background activity. Once these three criteria have been met, then the SP is accepted as a reasonable substitute for the REP, and Stage 1 is complete.

A Population Target

For validation we prefer a population target that is expected to capture 50 ± 30% of all values. After inspecting the results from several individual simulations, a range of 40–70 μg/ml plasma was arbitrarily selected to represent the therapeutic drug level range. The population target is therefore the fraction of the population that is within that range. Three times of interest were selected: 1, 3 and 6 hours after a dose at oscillatory steady state (OSS). We determined that OSS would be reached for all individuals by the 25th dose. Thus, we focused on drug levels obtained following the 30th dose. The fraction within the therapeutic range at the three times was calculated. Early in Stage 2 the Forecaster uses the SP, follows the above procedure and predicts a value for the population target at each of the three times.

The Approximate Prediction Interval

If one obtained a new test sample, repeated the above steps to obtain a new population target prediction, and repeated the entire process many times, then one would have sufficient data to calculate a confidence interval that measures the precision for the original population target prediction. The Forecaster uses a bootstrap resampling strategy (13) to mimic this sequence and to construct an approximate "confidence interval" for the above population target predictions (Fig. 1B). To avoid confusion with the traditional use of confidence interval, we refer to the interval described below as an API. Approximate is used as an adjective because the SP is not intended to replicate the true population. Any API, e.g., 50, 80, 95%, etc. may be calculated. For this validation, all APIs are 85%. We specify a low and high value for the population target such that, under ideal conditions, these values bracket the actual value for the REP approximately 85% of the time. To calculate an API, the experimental test sample can be resampled—bootstrapped—*with* replacement to give alias test samples (some sets are likely to be repeated). Alternatively, the SP can be resampled *without* replacement. In either case each bootstrapped sample is evaluated by the Forecaster, with no pre-conditions, as if it is a new test sample. We use the latter approach in this paper.

To specify an 85% API, 200 alias test samples are obtained from the SP. Each is analyzed separately by the Forecaster to give 200 alias SPs, and the latter are used to estimate a unique value of the population target. The estimates are rank-ordered, and the highest and lowest values are discarded such that the smallest interval containing 85% of the 200 is specified, thus completing Stage 2. Ideally (14), the lowest and highest 15 values are discarded. The 170 remaining values form the 85% API. The value of the population target calculated using the original experimental test sample will be within this range. Clearly, the actual population target value for the REP is also expected to be within the range.

An API will be more precise when the number of bootstraps is larger. Ideally, one would pool data from a thousand or more bootstraps to determine the API. Several independent test samples were evaluated by the Forecaster and then bootstrapped up to 600 times. We decided to terminate the bootstrap process once an API value decreased by <2% of the preceding

value. From the results we established that 200 would be an adequate number of bootstraps for this validation.

The Final Validation Task

To successfully complete the validation process it is necessary to answer a final question: Upon independent analysis of additional test samples (e.g., a total of 300), does one get an 85% API range that includes the known value of the REP population target, approximately 85% of the time? Each of 300 test samples (Fig. 2B) were taken from the REP and separately processed by the Forecaster. Each time a sample-specific SP was generated, a population target was calculated, and each SP was bootstrapped 200 times. Finally, 300 test sample-specific 85% APIs were specified. The information thus generated was sufficient to answer the final validation question and complete Part II. On average, and in the absence of bias, an 85% API is expected to include the true value for approximately 85 out of 100 identical size, independent random test samples taken from the REP. Although the SP represents the REP, its characteristics, in terms of Table I relationships, are not intended to be identical to those of the REP. They will be different. Thus, an 85% API can not capture the population target for the REP *exactly* 85% of the time. How much flexibility to allow for validation purposes is difficult to say given the absence of experience with such systems. We therefore arbitrarily decided that an API that captures the REP between 82 and 88% of the time will be acceptable.

RESULTS

The "Reference" Data Set

The population characteristics of the 1500-member "Reference" data set (Fig. 4) clearly reflect the actual PK data that served as template. The specified correlation between three pairs of parameters gave rise to significant, indirect correlation between three additional pairs of parameters (Table III). To provide OSS reference data each set of parameter values was used to calculate plasma levels at 1, 3 and 6 hours after dosing at OSS for the above dose regimen. The summary statistics of that data are listed in Table IV.

The REP serves only as a realistic *reference standard* for the validation, and is not expected to exactly match or reproduce the characteristics of the "Reference" data. A comparison between the population characteristics of a PK parameter in the "Reference" data set with its counterpart in the REP (Figs. 4 and 5) reveals differences and similarities. An expected difference is the slightly larger range for parameters in the REP. The correlations within the "Reference" data set are also present in the REP (Table III). In addition, a significant correlation exists between k_{01} and four other parameters, within the REP, whereas no such correlation exists within the "Reference" data set. It presumably results from a combination of the added individual PK uncertainty (see Appendix), the nature of the PK model, the choice of sampling times, and the fitting methodology. These factors cause an increase in parameter variance within the REP, except t_{lag}, relative to that within the "Reference" data set. As is evident in Fig. 5 for k_{12} vs k_{10}, there is also an increase in the conditional variance for correlated parameter pairs from the REP relative to the "Reference" data set. Even

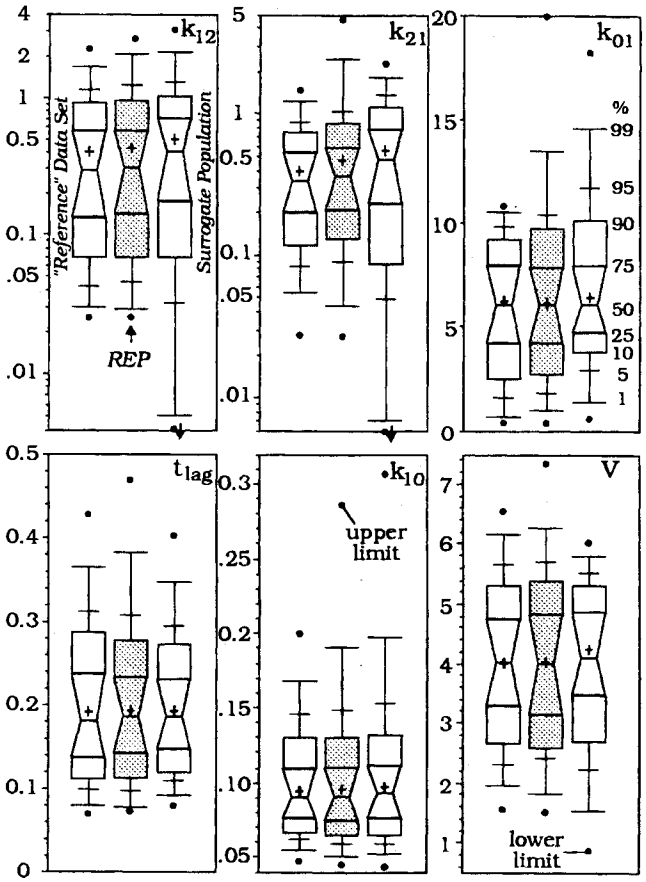

Fig. 4. Parameter Values. The unshaded bars and whiskers (left) characterize each of the six PK model parameters across the "Reference" data set. The lightly shaded bars and whiskers (middle) characterize each parameter within the REP. The unshaded bars and whiskers at right characterize each parameter within the SP. A cross marks each mean. ●: Either an upper or lower limit for all 1500 values. Some scales are logarithmic.

though such differences were seen when comparing parameter sets, it is particularly noteworthy that they were not evident in the corresponding sets of plasma levels at OSS (Table IV).

The Surrogate Population

Table II summarizes the experimental *test sample*. The Forecaster suggested that this data could have come from a probability density function (see Appendix) specified by Eq. 1, where θ is the vector of all parameter sets. The subscript s denotes that the function represents a SP.

$$p(\theta)_s = \pi(V) \cdot \pi(k_{01}|V) \cdot \pi(k_{10}|V) \cdot \pi(k_{12}|k_{10}) \cdot \pi(k_{21}|k_{12}) \cdot \pi(t_{lag}) \quad (1)$$

The parameters V and t_{lag} were judged independent. Therefore, the values of V in Table II could represent a random sample taken from a marginal univariate density, $\pi(V)$. "Normal" was selected from Fig. 3 to represent $\pi(V)$ and "slightly skewed to larger values" was selected to represent $\pi(t_{lag})$. There are four conditional terms. As an example, $\pi(k_{01}|V)$ is the conditional density function for k_{01} given a value of V. The Forecaster suggested that the Table II values could represent a random

Table III. Correlation[a] Between Parameter Pairs Within Populations

Parameter[c]	V	k_{01}	k_{10}	k_{12}	k_{21}
k_{01}	<0.0001[b]				
	0.0033				
	none				
k_{10}	<0.0001	<0.0001[d]			
	<0.0001	<0.0001			
	<0.0001	none			
k_{12}	<0.0001[d]	<0.0001[d]	<0.0001		
	<0.0001	<0.0001	<0.0001		
	<0.0001[d]	none	<0.0001		
k_{21}	<0.0001	<0.0001[d]	<0.0001[d]	<0.0001	
	<0.0001	<0.0001	<0.0001	<0.0001	
	<0.0001	none	<0.0001[d]	<0.0001	
t_{lag}	none	none	none	none	none

[a] The Kendall Rank Correlation test is used to test for correlation between parameter pairs within a data set.
[b] Within a cell the correlation values given, from top to bottom, correspond to the SP, the Reference Experimental Population, and the "Reference" data set. A correlation is treated as real when $p \leq 0.05$, and in these cases the p value is given. Otherwise no correlation (none) is assumed.
[c] Units are as in Table II.
[d] The observed correlation between this pair of parameters is the result of indirect correlation.

sample taken from $\pi(k_{01}|V)$. The Forecaster used Table I menu choices to characterize each of the six terms in Eq. 1 such that a random sample of 100 from the selected form of a term had a mean and variance that were as close as possible to the corresponding values in the test sample. The Table I regression function choices were $\kappa = 4$, 6, 3 and 2 for the second through fifth terms, respectively, in Eq. 1. The conditional variance choices were "proportional to predicted," for the second, fourth and fifth terms, and "constant" for the third term. The conditional density function choices were $\kappa = 3$ for $\pi(k_{01}|V)$, $\kappa = 8$ for $\pi(k_{10}|V)$, $\kappa = 4$ for $\pi(k_{12}|k_{10})$, and $\kappa = 6$ for $\pi(k_{21}|k_{12})$. Subsequently, it is this fully characterized form of Eq. 1 that is specified when we cite Eq. 1.

Table IV. Descriptive Statistics for the 1500 Drug Plasma Levels at Three Times After Dosing at Oscillatory Steady State (OSS)

	Mean	Std. Dev.	CV(%)	Percentile 1	10	50	90	99
			1 hour					
Surrogate	68.1	13.7	20.1	40.6	50.9	67.0	86.7	101.4
REP[a]	71.1	13.3	18.8	39.3	53.9	71.6	90.6	99.2
"Reference"[b]	70.8	13.1	18.6	42.3	53.9	70.9	88.5	100.5
			3 hours					
Surrogate	53.4	11.9	22.3	29.2	38.4	52.7	69.6	79.3
REP	56.3	11.1	19.6	31.5	41.8	57.4	71.8	78.6
"Reference"	53.4	11.9	22.2	32.9	38.3	52.7	69.5	79.8
			6 hours					
Surrogate	44.2	9.9	22.3	23.9	32.1	43.6	57.8	65.9
REP	46.1	9.1	19.8	25.9	34.3	47.1	59.1	64.7
"Reference"	46.0	8.9	19.3	26.9	34.2	46.2	57.9	66.3

[a] The Reference Experimental Population.
[b] The "Reference" Population.

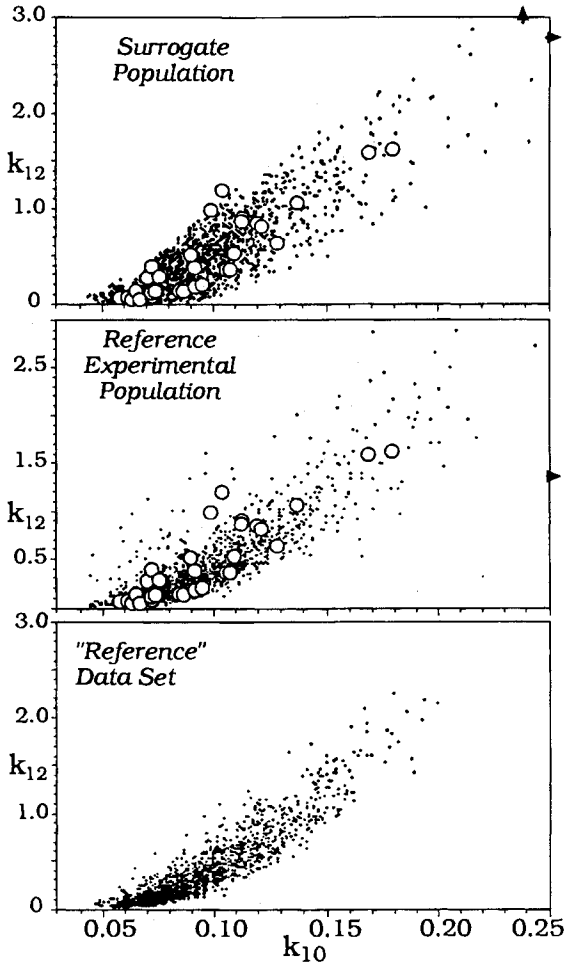

Fig. 5. Correlated Parameter Values. The correlation reported in Table III between different pairs of parameters is illustrated for k_{12} and k_{10}. The open circles are paired values from the 30-member test sample (Table II).

A set of parameter values to represent one hypothetical individual is assembled by randomly sampling one value from each term in Eq. 1 in sequence from left to right. In this way we assembled 1500 sets to represent the SP. The expectation is that test samples taken from the SP and the REP will be statistically indistinguishable (Smirnov test, $p > 0.05$). Indeed each of 100 random (alternate) test samples from Eq. 1 met that condition (not shown). Also, significant relationships were present within the SP for the same nine correlated pairs that were identified within the REP (Table III). Four of these pairs are specified by Eq. 1. The other five are the result of indirect correlations. In some cases the pattern of the relationship within a parameter pair is visibly different between the SP and the REP. As an example, contrast the relationship between k_{12} and k_{10} in Fig. 5 within the SP and the REP. Although differences are evident at the two dimensional level, because of the large interindividual variance, test samples taken from each population are statistically indistinguishable (Smirnov test, $p > 0.05$). Very large samples will, however, be distinguishable.

Drug plasma levels at 1, 3 and 6 hours after dosing at OSS were predicted using the SP, and in Fig. 6 are contrasted to the values calculated from the REP. The descriptive statistics

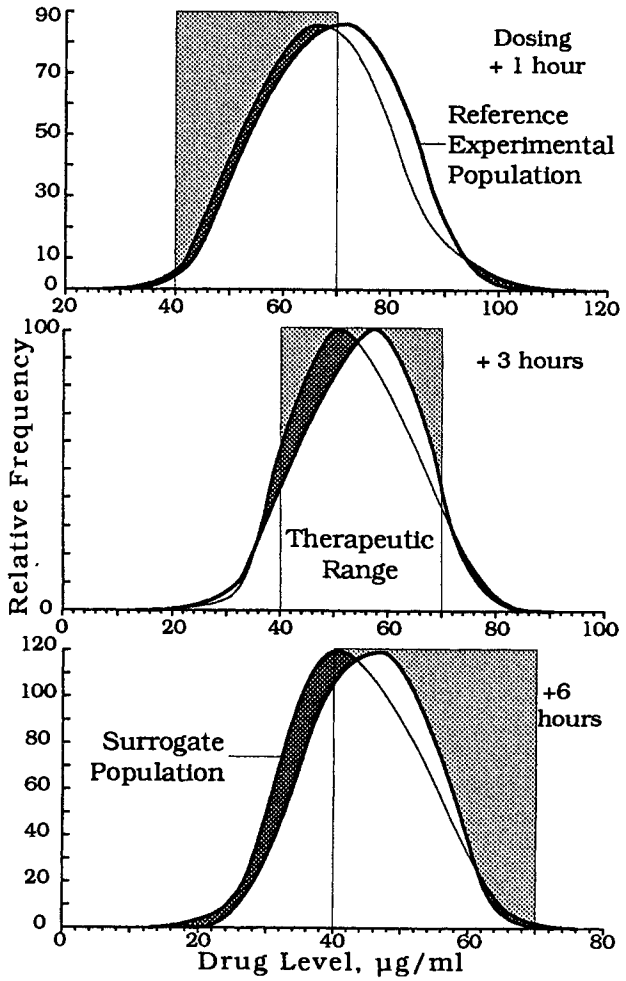

Fig. 6. Drug Level Frequency Distributions. Each curve is a smoothed frequency distribution of 1500 drug plasma levels (2.5 μg/ml intervals) calculated for the conditions specified in the text for the REP and the SP at 1, 3 and 6 hours after dosing at OSS. The rectangle indicates the population target (therapeutic range).

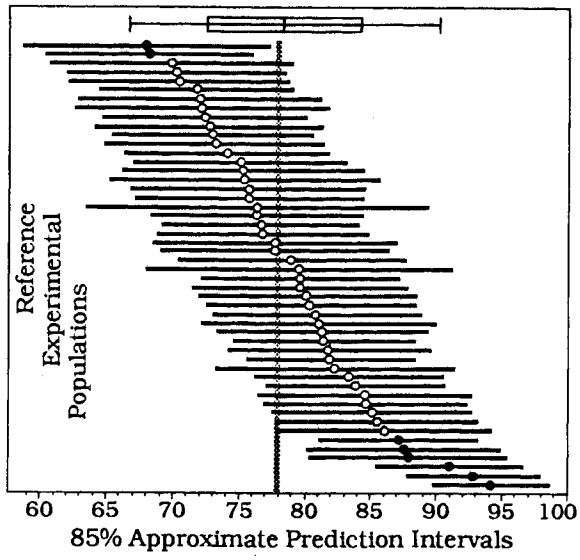

Fig. 7. Approximate Prediction Intervals. Each horizontal bar gives an 85% API for the percent of the REP that is expected to be within the therapeutic range, 40–70 μg/ml, for the conditions specified in the text. Each API is calculated for a SP that was independently generated by the Forecaster when given one of 300 independent 30-member test samples taken randomly from the REP (Fig. 2B); 50 of these, selected arbitrarily, are shown ordered by median value. The vertical bar marks the actual population target value for the REP, 78% (specifically, 77.9%). ○: The midpoint of each API that captures the REP target value. ●: The midpoint of each API's that falls to capture the REP target value. 84.0% of the 300 API (252) captured the population target value in the REP. The bar and whiskers at the top designate the mean of 300 API medians ±1 and 2 SD. Of the 300 API's the one that missed most on the low side was the interval 49 to 67%, and the one that missed most on the high side was 90 to 99%.

are in Table IV. Clearly, the SP makes a reasonable substitute for the REP in terms of providing projected drug levels across the population, even at OSS where the independent variable settings are different from those under which the original data was collected. At one hour 56% of the REP is predicted to be within the therapeutic range, whereas the actual value is 45%. Similarly, three hours after dosing 77% is predicted to be within the therapeutic range, whereas the actual value is 78%. At six hours, 64% is predicted; the actual value of is 73%. How reliable—how accurate—are these predictions?

To obtain a measure of reliability 200 alias population target values were calculated. The 15 smallest and largest values were discarded. The remaining values formed the 85% API. At one hour after dosing between 41 and 59% of the REP is predicted to be within the therapeutic range with a probability of approximately 85% of being correct (15% of being wrong); the actual value is 45%. At three hours between 70 and 84% is predicted to be within the therapeutic range, whereas the actual value is 78%. Finally, at six hours between 60 and 76% is predicted to be within the therapeutic range, and the actual value is 73%.

Validation of Predictions

How frequently will an 85% API capture the actual percentage of the REP that is within the therapeutic range? To answer this question 300 experimental test samples from the REP were separately used to calculate an 85% API for the population target at 3 hours after dosing at OSS. Of these 252—84%—captured the actual known population target value of 78%. Fifty, selected at random, of these API are represented in Fig. 7. For a 95% API, the intervals in Fig. 7 will be wider. This completes Part II and successfully concludes the validation.

DISCUSSION

The Validation

For the boundary conditions and assumptions specified, the prototype Forecaster makes useful population predictions, and the API functions as designed by providing a useful measure of prediction reliability. The results support the proposition that the hypothetical individuals that comprise a SP can be used to predict clinical trial outcomes and to answer a variety of "what if" questions. Clearly, further testing and validation is needed, especially using population targets that are PD-based, and they are in-progress. Extension of this approach to include PD and covariate data may allow one to forecast results with limited or non-existent current data.

The Surrogate Population Approach

Among other things, a SP provides a useful representation of the interindividual variance structure of a population, and therefore can provide a valued resource. One could use it to address sample size questions, to query the merits (for example) of shifting to a sustained release dosage form or to an alternate route of delivery, and of using optional clinical trial designs. It provides the means to bring into sharper focus both what is known and what remains unknown (15). The expectation is that a SP can be used to increase the probability that, during the clinical phase of drug development, primarily the pivotal and supportive studies will be conducted. Arguments have already been made that when this is done, drug development will be concluded sooner, and at a lower cost (16–19).

To have sufficient flexibility to address development research issues, our approach is that the SP needs to allow generation of the individual output data that provides the basis for decision making. In designing the prototype around the conditional distribution approach (12) we hypothesized that there is now sufficient knowledge so that one can *adequately* represent interindividual relationships between inter-related variates across a population using selections made from a limited number of options (e.g., Table I). Because data from humans can have large variances, we can reduce the number of menu options to a manageable number. The number of menus, the number of options within each and the exact options used are open issues. The characterization of the SP, such as that given by Eq. 1, is not intended to be the an optimum or most likely representation of the actual characteristics of an experimental population. That is not our objective.

A successful and useful drug development decision support system will generate predictions for different decision paths that can be weighed against each other. Providing a prediction in the form of an approximate prediction interval (API), rather than an expected population mean drug level or response value, provides the needed weight, and so will be more useful in a critical decision context (4).

Flexibility and Prior Knowledge

Success in meeting the stated validation challenge derives from system flexibility and use of prior knowledge. Prior knowledge from five sources is added to the experimental data. First is knowledge that supports the assumption that the marginal distribution for each model parameter is smooth and adequately representable by one of the Fig. 3 functions. These functions add tails to the experimental data. Next, is the knowledge that all biologically realistic parameter values lie in the positive quadrant of all possible values. Third, significant apparent correlation between parameter pairs in a test sample is taken to be real and so additional information is added by deciding that a Table I function can adequately represent the relationship. Fourth is the knowledge that all clinically relevant PK/PD parameters and covariates will have finite limits. Finally, the conditional variance for a correlated parameter pair is adequately represented by one of three options. Adding such information is supported by the large PK and PD literature that has accumulated over the past three decades. Unusual cases and exceptions are known, however. There is clearly a risk that some portion of this added information will be inappropriate for the case at hand. Therefore, seeking evidence that might contradict a questionable assumption can be included among the experimental design objectives.

Forecaster flexibility is also a consequence of how we have implemented the conditional distribution approach. A description of a six dimensional parameter space can take many forms. Equation 1 is just one of 720 (6!) possible combinations of marginal and/or bivariate conditional terms that were actually considered before Eq. 1 was specified. During Stage 2, any one bootstrapped sample may be represented by any one of these 720 different versions of the population joint density function. In fact, many of these versions were actually selected and then used in the process of constructing 300 85% API (Fig. 7).

Flexibility also comes from having a finite menu of options to represent each term in Eq. 1. A marginal term, such as $\pi(V)$, will be represented by one of nine different density functions, yet each of these will have the same mean and variance. Their distribution patterns and ranges will differ. Each conditional distribution will also be represented by one of these nine functions. In addition there are three conditional variance options, and the regression relationship will take any one of eight forms. Thus, each of the four conditional terms in Eq. 1 (as written) can be represented in one of $9 \cdot 8 \cdot 3 = 216$ different ways. It follows that Eq. 1 can take any one of $(2 \cdot 9 + 4 \cdot 216) = 882$ different characterizations. For the same independent variable conditions, the range and pattern of drug levels across the population that are predicted from several of the 882 characterizations can be quite different, even though there may be little difference in the predicted population mean levels.

The approach needs deterministic models but is flexible about their nature. Equation 1 can be based on any of a variety of model types where model parameter values are independently obtained for, and assigned to, an individual. We used a classical compartmental model (7,20) because that was the approach selected by the team that produced the template data. Alternatively, one can use a noncompartmental approach or physiological PK models (21), or even an infrequently used modeling approach (e.g., 22–24). Equation 1 can be easily extended to include PD model parameters and concomitant variables, either discrete or continuous.

Forecaster Reliability

Is the API 85% as stipulated? Is an assessment using 300 test samples sufficient to answer the question? The answer depends on the level of accuracy sought and on the stringency of the criteria used. The 85% API is intended to provide a measure of reliability for a single population target forecast, presumably at a time when actual development targets are being set. Therefore, for this validation we aimed only to confirm that the prediction interval is *approximately* 85%. We specified that having the known population target value fall within the API between 82 and 88% of the time (246 to 264 times out of 300) would be acceptable. Having an accuracy of $85 \pm 5\%$ may also be acceptable. As the SP approach evolves, the issues of API width and accuracy will need more research. Forecaster design and the validation strategy are built upon numerous assumptions that are supported by several literature sources. It follows that the accuracy of an API may be as dependent on the appropriateness of these built-in assumptions as on Forecaster sample sizes, or the exact characterization selected for the SP.

Questions

Taken together the test results provide the feasibility support sought. Additional research, however, is needed to establish the strengths and weaknesses of the SP approach. Caution is also needed. Even though the REP is realistic and based on real data, it had to be generated. Validation is still needed that uses existing early PK/PD data to predict later (also existing) PK/PD data.

Both the testing and the results raise many interesting questions. What constitutes adequate testing and validation for such a complex decision support system? Have the validation criteria and plan been sufficiently exacting? What is the relationship between experimental test sample size and the width of an API? Within the context of drug development, is an 85% API appropriate or too demanding? Are other API, 80% or 90%, also reasonable? To generate the API, is it better to resample with replacement from the experimental data or, as done here, resample without replacement from the SP? Can the Forecaster design accommodate missing data and incomplete experimental parameter sets? Most people are better able to understand complex information when they see it displayed. What is the best strategy to visualize the important population information features within a variety of multidimensional clinical outcomes? These and other questions are being addressed as part of our ongoing research.

APPENDIX

Generating The Reference Experimental Population

The template data (1) consists of drug plasma levels—not body weight adjusted—measured at 20 times (0.083 to 48 hours) following a single 200 mg tablet dose. Individual data were fit to the model using PCNonlin (SCI, Cary, NC) assuming a Poisson error model with a positive parameter constraint. Had the team fit the original data to a physiologic PK model (21), the Forecaster strategy would have been unaltered, but Eq. 1 would be different.

The template data is used to generate a *reference standard* assuming that each PK parameter comes from a population characterized by a smooth, unimodal, distribution. When correlations are seen (real or apparent), we assume that the relationship is continuous in the parameter ranges of interest, and that the actual underlying relationship can be represented by one of several regression functions, such as in Table I. There is no compelling reason to assume that the variance about the regression is constant or that the residuals are normally distributed. To allow for many of the varied relationships reported in the literature, we assume that the conditional distribution of one parameter, given a value of another, can be reasonably represented by one of several density functions, such as in Fig. 3.

Desiring a realistic *reference standard*, we arbitrarily specified that three significant correlations should exist, as detailed below, and that the remaining three parameters, V, k_{01} and t_{lag}, would be independent. In the latter three cases, a density function was selected at random from Fig. 3 to describe each. The density function selected for V was scaled such that the mean and variance of 100 random values taken from it were identical to the corresponding template data values. We repeated this procedure for k_{01} and for t_{lag}. For each pair of correlated parameters we arbitrarily specified a regression function form Table I to represent the relationship. We randomly selected a Fig. 3 option to represent the conditional distribution. The conditional variance was specified to be either constant or proportional to the expected parameter value. Using k_{10} given a value of V, $(k_{10}|V)$, as an example, we adjusted these three properties of the relationship until the mean and variance of 100 values of k_{10} taken at random from the adjusted conditional density of $(k_{10}|V)$ were as close as the constraints allowed to both the mean and variance of the k_{10} template data. The same protocol was followed for $(k_{12}|k_{10})$ and finally for $(k_{21}|k_{12})$. Thus, a unique population of PK parameter sets was specified. A single random sample was then taken from each of these six characterizations—in sequence—to generate a set of single values of V, k_{01}, t_{lag}, and then k_{10}, k_{12} followed by k_{21}, and that set represented the first hypothetical individual in the "Reference" data set. Additional random samples were taken until the "Reference" data set (Fig. 2) totaled 1500. The selected Table I and variance choices remained unknown for the duration of the validation.

Four more steps are needed to get the Reference Experimental Population (REP). First, each "Reference" set is introduced into the PK model at the same independent variable settings that gave rise to the template data, to generate error-free drug level data, C_R (20 times, 0.083–48 hours). Random error, selected at random from a normal density having a 10% CV (11), is then added separately to each C_R to obtain simulated experimental drug levels, $C_E = C_R + \epsilon$, where the variance of ϵ is proportional to C_R. Using the WinNonlin 1.1 (SCI, Cary, NC) Gauss-Newton weighted least squares fitting algorithm (20) we fit each set of C_E to the PK model. To minimize the type of post-fitting problems discussed by Purves (10) we added the constraint that all parameter estimates be within the following range limits: V: 1–10 liters; k_{01}: 0.1–10 hr^{-1}; k_{10}: 0.03–0.3 hr^{-1}; k_{12}: 0.02–3.0 hr^{-1}; k_{21}: 0.02–3.0 hr^{-1}; and t_{lag}: $\leq$ 0.5 hrs.

Forecaster Design and Execution

The development and explanation of the Forecaster detailed in (1) is summarized here. Presume that each of n sets of experimental data have been individually and adequately fit to a common PK model and to a corresponding common PD model, Eq. 2.

$$\mathrm{E}[C] = f_1(x, \theta) \text{ and } \mathrm{E}[E] = f_2(x, \theta) \qquad (2)$$

The function f_1 describes a PK model of the type specified in Methods; the function f_2 describes the drug's pharmacodynamics at settings of the same independent variables, x, and for the vector of model parameter values θ. Assume that the uncertainty in the individual parameter estimates is small relative to the population variance, and that all model parameters are treated as random variates. The mathematical and statistical algorithms needed to characterize θ come from the IMSL® Math/Library® and Stat/Library® (Visual Numerics, Inc., Houston, TX). The Forecaster methodology is programmed as a series of ten steps.

1. Randomly order all p variates within each category (PK, PD, etc.) and factorize their joint density as a product of marginal and conditional density functions. Let $\pi(\beta_k)$ stand for the marginal distribution of kth variate. Let $\pi(\beta_k|\beta_1, \beta_2, \ldots,$

β_{k-1}) stand for the conditional distribution of β_k, k $\geq$ 2, given specific values of the other variates.

$$p(\theta) = \pi(\beta_1) * \pi(\beta_2|\beta_1) * \pi(\beta_3|\beta_1, \beta_2) * \cdots * \pi(\beta_p|\beta_1, \beta_2, \ldots, \beta_{p-1}) \quad (3)$$

Presuppose that sufficient prior knowledge is available to narrow the range of choices needed to characterize each marginal and conditional distribution to the 9 univariate density functions in Fig. 3, the 8 bivariate regression functions in Table I, and 3 conditional variance options.

2. a) Starting with β_1, select M (here, $M = 100$) numbers from a Fig. 3 option and scale them to have the same mean and a variance as the experimental values. Refer to them as β_{1_j}, where j refers to the density function. Use the Smirnov test, to compare each β_{1_j} with the experimental β_1, and refer to the computed test statistic as a *matching index* (*MI*). A *MI* approaching 1.0 is preferred. b) Repeat Step 2a for the remaining density functions, and then select the set with the largest *MI* as best representing β_1.

3. Reduce $\pi(\beta_k|\beta_1, \beta_2, \ldots, \beta_{k-1})$ to a distribution conditioned on just one other variate, i.e.

$$\pi(\beta_k|\beta_1, \beta_2, \ldots, \beta_{k-1}) \cong \pi(\beta_k|\beta_1 \text{ or } \beta_2 \text{ or } \cdots \text{ or } \beta_{k-1}) \cong \pi(\beta_k|\beta_{1*}) \quad (4)$$

without loss of generality. Condition β_k on only β_1 or β_2 or β_{k-1}, whichever exhibits the largest correlation with β_k. Evaluate each possible conditional distribution in Eq. 3, in sequence, for evidence of correlation by calculating a Kendall rank correlation coefficient for each pair. Disregard any correlation having $p > 0.05$ (adjustable). Rank-order the remaining coefficients. Designate the variate exhibiting the largest correlation with β_k as β_{1*}. If no significant correlations are evident then Eq. 4 reduces to $\pi(\beta_k)$. The general form of Eq. 3 thus becomes Eq. 5.

$$p(\theta) \cong \pi(\beta_1) * \pi(\beta_2|\beta_1) * \pi(\beta_3|\beta_1 \text{ or } \beta_2) * \cdots * \pi(\beta_p|\beta_1 \text{ or } \beta_2 \text{ or } \cdots \beta_{p-1}) \quad (5)$$

4. Let the expected value of β_k given a value of β_{1*} be $E[\beta_k|\beta_{1*}] = h_\kappa(\beta_{1*})$, where h_κ designates a Table I regression functions.

5. Assume that the conditional variance for all β_k given β_{1*} is either constant, proportional to (the expected value of) β_k, or proportional to β_k^2. Use prior knowledge or the residuals generated during Step 4 to select one of these three options.

6. a) Identify one Fig. 3 option to represent $\pi(\beta_k|\beta_{1*})$. For each value of β_{1*} simulate a corresponding random value of β_k: simulate M numbers from a density function and transform them such that they have the conditional mean in Step 4 and the conditional variance in Step 5. Compute the *MI* between the experimental and simulated values of β_k. b) Repeat Step 6a 20 times and compute an *average matching index, AMI.*

7. a) For each additional Fig. 3 option repeat Steps 5 and 6. Select the one distribution having the largest *AMI* as best representing β_k. b) For an additional conditional term, repeat Steps 3 through 7a.

8. Repeat steps 5 through 7 for all other variates. Each of M sets is then introduced into Eq. 2 to yield sets of individual outcome values. At each time use the Smirnov test to compare these M values and the corresponding n experimental values. If no significant difference is detected, then accept $p(\theta)$ and designate it $p(\theta)_S$. Otherwise return to Step 1.

9. A large subset of $p(\theta)_S$ is used to represent the *Surrogate Population.* To generate a SP take Z (typically, $Z \geq 1{,}500$) random samples from $p(\theta)_S$ and introduce each into the outcome functions of interest, e.g., Eq. 2, and calculate a value of the *population target.*

10. To measure the precision or reliability of the *population target* prediction from step 9, select one of two options: i) the n sets of the *experimental* test sample are resampled *with* replacement (13) and each time Steps 2 through 8 are followed to obtain an alias description of $p(\theta)_S$; ii) the SP from step 9 is resampled *without* replacement, and each time Steps 2 through 8 are followed to obtain an alias description of $p(\theta)_S$. Then, i) or ii) is repeated L times, (e.g., $L = 200$). Finally, the L density descriptions are processed according to step 9 to give L alias values of the *population target.* Use a procedure to calculate an approximate $1 - \alpha$ level equal tail confidence interval (25) and use it as the Approximate Prediction Interval: it serves as an indication of the precision and a measure of the reliability of the *population target* prediction.

Data Used

All data used for this validation is available from the corresponding author.

ACKNOWLEDGMENTS

This work was supported in part by Scientific Consulting Inc., the University of California Biotechnology Program (SG), GM08388 from NIH (CAH), grants from the UCSF Academic Senate Committee on Research (CAH and SG), and patent royalties (CAH). Prior support was provided by Syntex Research, Inc. and grant N00014-91-J1455 from the ONR. We thank Peter Shih for technical contributions, Dan Weiner for constructive scientific and conceptual commentary, and Davide Verotta and Lisa Bero for helpful discussions and useful commentary on this manuscript.

REFERENCES

1. C. A. Hunt, S. Guzy, and D. L. Weiner. *Stat. Med.*, in press, due: (1997); and http://www.mis.ucsf.edu/decision/pub/forecast/
2. L. B. Sheiner and T. M. Ludden. *Ann. Rev. Pharmacol. Toxicol.* **32**:185–209, 1992.
3. N. G. Best, K. K. C. Tan, W. R. Gillas, and D. J. Spiegelhalter. *J. Pharmacok. Biopharm.* **23**:407–435, 1995.
4. R. A. Howard and J. E. Matheson. *The Principles and Applications of Decision Analysis, Vol. I.* Menlo Park, Strategic Decisions Group, 1989.
5. R. M. Sailors, T. D. East, C. J. Wallace, D. A. Carlson, M. A. Franklin, L. K. Heermann, A. T. Kinder, R. L. Bradshaw, A. G. Randolph, and A. H. Morris. *Proc. AMIA Annu. Fall Symp.* **2**:234–238, 1996.
6. F. E. Harrell Jr, K. L. Lee, and D. B. Mark. *Stat. Med.* **15**:361–387, 1996.
7. J. G. Wagner. *Pharmacokinetics For The Pharmaceutical Sciences.* Technomic Publication Co., Lancaster, PA, USA, 1993.
8. T. J. Woodruff, F. Y. Bois, D. Auslander, and R. C. Spear. *Risk Anal.* **12**:189–201, 1992.
9. K. Murata and K. Kohno. *Biopharm. Drug Dispos.* **10**:15–24, 1989.

10. R. D. Purves. *J. Pharmacokin. Biopharm.* **24**(1):79, 1996.
11. P. M. Laskarzewski, D. L. Weiner, and L. Ott. *J. Pharmacok. Biopharm.* **10**:317–334, 1982.
12. M. E. Johnson. *Multivariate Statistical Simulation*. New York, John Wiley and Sons, pp. 43–48, 1987.
13. B. Efron and R. J. Tibshirani. *Monograph on Statistics and Applied Probability, No. 57. An Introduction to the Bootstrap*. New York, Chapman and Hall, 1993.
14. V. K. Roltatgi. *Statistical Inference*. New York, John Wiley, pp. 616–617, 1984.
15. N. H. G. Holford. *Clin. Pharmacokin.* **29**:287–297, 1995.
16. J.-L. Steimer, M.-E. Eblin, and J. Van Bree. *Eur. J. Drug Metab. Pharmacokinet.* **18**:61–76, 1993.
17. C. C. Peck. Population approach in pharmacokinetics and pharmacodynamics: FDA view. In: *New Strategies In Drug Development and Clinical Evaluation*, M. Rowland, L. Aarons, eds. Luxembourg, Commission of the European Communities, pp. 157–168, 1992.
18. C. C. Peck et al., *Clin. Pharmacol. Therap.* **51**:456–473, 1992.
19. M. Hale, W. R. Gillespie, S. K. Gupta, B. Tuk, and N. H. G. Holford. *App. Clin. Trials* **5**:35–40, 1996.
20. J. Gabrielsson and D. L. Weiner. *Pharmacokinetic And Pharmacodynamic Data Analysis: Concepts and Applications*. Upsula, Swedish Pharmaceutical Press, 1994.
21. W. J. Jusko. Guidelines for the collection and analysis of pharmacokinetic data. In: *Applied Pharmacokinetics, Principles of Therapeutic Drug Monitoring*, Third Ed., W. E. Evans, J. J. Schentag and W. J. Jusko, eds. Vancouver, WA, Applied Therapeutics, pp. 2.1–2.32, 1992.
22. J. L. Matis and T. E. Wehrly. *J. Pharmacok. Biopharm.* **18**:589–607, 1990.
23. P. Macheras. *Pharmac. Res.* **13**:663–670, 1996.
24. W. A. Colburn. *J. Pharmacok. Biopharm.* **11**:389–400, 1983.
25. M. E. Johnson. *Multivariate Statistical Simulation*. New York, John Wiley and Sons, pp. 160–162, 1987.