Pharmaceutical Research, Vol. 22, No. 4, April 2005 (© 2005)
10.1007/s11095-005-2499-5

Research Paper

Data Supplementation: A Pharmacokinetic/Pharmacodynamic Knowledge Creation Approach for Characterizing an Unexplored Region of the Response Surface

Ene I. Ette,[1,2] Hui-May Chu,[1] and Christopher J. Godfrey[1]

Received April 12, 2004; accepted June 21, 2004

Purpose. To develop a data supplementation [i.e., a pharmacokinetic/pharmacodynamics (PK/PD) knowledge creation] approach for generating supplemental data to be used in characterizing a targeted unexplored segment of the response surface.
Methods. The procedure for data supplementation can be summarized as follows: 1) statement of the objective of data supplementation for PK/PD knowledge creation, 2) performance of PK knowledge discovery, 3) PK data synthesis for target dose group(s), 4) covariate data synthesis for virtual subjects in the target dose group(s), 5) discovery of hidden knowledge from real data set to which supplemental data will be added, 6) implementation of a data supplementation methodology, and 7) discovery and communication of the created knowledge. A nonparametric approximate Bayesian multiple supplementation and its modification, structure-based multiple supplementation, which is an adaptation of the approximate Bayesian bootstrap, is proposed as a method of data supplementation for PK/PD knowledge creation. The structured-based multiple supplementation methodology was applied to characterize the effect of a target dose of 100 mg that was unexplored in a previously concluded study that investigated the effect of 200- and 600-mg doses on biomarker response.
Results. The target dose of 100 mg was found to produce a response comparable with that of the 200 mg and better than that obtained with the 600 mg.
Conclusions. Implementation of the PK/PD knowledge creation process through data supplementation resulted in gaining knowledge about a targeted region of a response surface (i.e., the effect of a target dose) that was not previously studied in a completed study without expending resources in conducting a new study.

KEY WORDS: data supplementation; multiple supplementation; PK/PD knowledge creation; structure-based multiple supplementation; nonparametric approximate Bayesian bootstrap.

INTRODUCTION

Pharmacokinetic/pharmacodynamics (PK/PD) knowledge creation is the process of building upon current understanding of data that is already acquired by generating more data (information) that can be translated into knowledge. It entails the use of (valid) models to synthesize data, estimate inestimable uncertainty, or supplement data for further knowledge acquisition.

When there is a considerable amount of information about the drug, synthesizing data into a coherent package that indicates the drug developer has understanding of the pharmacology and, eventually, good control over the therapeutics of the drug provides a means for knowledge creation about the drug being developed. Data synthesis is performed when available knowledge about the drug is used to simulate a clinical trial to explore study outcome when various controllable and uncontrollable factors are varied. This is a knowledge creation process because the objective is to obtain knowledge about the unknown (i.e., unexplored region of the response surface) using valid models. A case in point is the use of clinical trial simulation to investigate the exposure-response relationship in a first time-in-man study. This involves not only extrapolation of PK/PD from animal to man, but also the exploration of the response surface, hitherto unknown, for a new compound about to be introduced into man (1). Data synthesis via clinical trial simulation offers the means of generating complex data sets, which may include the influence of prognostic factors, sample size, and dropouts, for testing new competing analysis methods (2,3).

Physiologically based pharmacokinetic (PBPK) modeling is a modeling approach that lends itself to knowledge creation (1). The result is a model that predicts the qualitative behavior of the experimental time course without being based on it. Refinement of the model to incorporate additional insights gained from comparison with experimental data yields a model that can be used for quantitative extrapolation beyond the range of experimental conditions. That is, the model allows predictions to be made of the kinetic behavior of drug at various dose levels and routes of administration.

[1] Department of Clinical Pharmacology, Vertex Pharmaceuticals, Inc., Cambridge, Massachusetts 02139, USA.
[2] To whom correspondence should be addressed. (e-mail: ene_ette@vrtx.com)

Parameter estimation without an appropriate assessment of reliability of the estimates yields no confidence in such estimates. Estimation of uncertainty enables the use of such parameter estimates in data synthesis. Embarking on data synthesis (e.g., clinical trial simulation) using model parameter estimates without associated uncertainty or poorly defined uncertainty will produce unreliable outcomes. Sometimes it is impossible to obtain standard errors for population model parameter estimates when small sample sizes are used for population PK/PD modeling. The bootstrap with winsorization has been proposed for the estimation of inestimable uncertainty—standard errors—for population PK/PD parameters that are usually not obtainable using software such as NONMEM because of small sample size (4).

Data supplementation deals with the use of models on available data to generate supplemental data that will be used to characterize a targeted unexplored segment of the response surface. The assumptions about models to be used in data supplementation are not as stringent as those required for data synthesis. That is, the use of predictive models is not an absolute necessity.

The intent of knowledge creation is the characterization of unexplored response surface to aid our understanding of drug action. The response surface can be described as three-dimensional. On one axis are the input variables (controllable factors) such as dosage regimen and concurrent therapies. Another axis incorporates patient characteristics, which summarizes all the important ways patients can differ that affect the benefit to toxic ratio (5). That is, the response surface describes the relationship between the therapy and the effects and how this relationship varies with patient characteristics and time to explain tolerance or sensitivity. For rational drug development and the optimization of individual therapy, this response surface must be mapped for the target population. This shift has occurred because of a concern for maximizing the benefit/risk ratio for individual patients in addition to answering the question of efficacy. Regarding knowledge of the response surface, PK/PD knowledge discovery (6) and creation, as described later, greatly improve the precision of this process, which in turn can result in rational drug development with optimized dosing strategies. PK/PD knowledge discovery is the nontrivial process identifying valid, novel, potentially useful, and ultimately understandable patterns in data by characterizing data structure by means of a model (6).

With the PK/PD knowledge discovery process, information (data) is turned into knowledge, and the PK/PD knowledge creation process results in more knowledge generation. Knowledge extracted or created from a clinical trial data set can then be used for decision making. Thus, after the completion of the PK/PD knowledge creation process, a better comprehension is gained about the response surface. This knowledge and comprehension makes wisdom for rational drug development possible, because wisdom (the knowledge and ability to make the right choices at the opportune time) is the final step of good mission-critical decision making. That is, the created knowledge can then be appropriately applied in the design and conduct of appropriate mission-related clinical trials, or the progression of a compound in development.

Knowledge creation is an emerging, interdisciplinary research field that lives at the intersection of computer science (database, artificial intelligence, graphics, and visualization), statistics, and an application domain such as clinical pharmacology in general, and pharmacometrics in particular.

In this paper we propose data supplementation as a PK/PD knowledge creation methodology for characterizing a targeted segment of an unexplored region of the response surface.

GENERAL STEPS IN THE PK/PD KNOWLEDGE CREATION PROCESS

PK/PD knowledge creation from a clinical trial data set is a process that can be formalized into a number of steps. In this section, we provide a general framework for the steps needed to be taken in the PK/PD knowledge creation process. These steps could vary depending on the type of knowledge creation approach involved. Subsequently, data supplementation—the PK/PD knowledge creation approach of focus in this paper—is discussed.

Briefly, the steps in the PK/PD knowledge creation process are as follows:

1. Statement of the objective of the PK/PD knowledge creation process;
2. A data set and/or a valid model summarizing the discovered knowledge from a prior PK/PD knowledge discovery process;
3. Performance of knowledge creation (i.e., data synthesis, estimation of inestimable uncertainty, or data supplementation);
4. Analysis of the data generated in step 3;
5. Application of the knowledge created; and
6. Communication of the created knowledge.

The objective of the PK/PD knowledge creation process must be clearly defined. With a clear objective in mind, the path chosen for the PK/PD knowledge creation process can be delineated. For PK/PD knowledge creation via data synthesis, valid models are needed. Data synthesis performed using clinical trial simulation requires the use of valid input/output models for the PK/PD knowledge creation process. Data supplementation, on the other hand, requires model assumptions that are not as stringent as the assumptions made when analyzing the data created by a data supplementation methodology.

Once data synthesis or supplementation is performed, the data must be analyzed for the created knowledge to be extracted. This can be performed using statistical or population PK/PD modeling approaches chosen by the pharmacometrician/pharmacokineticist. There will be variations in steps 3 and 4 of the PK/PD knowledge creation process depending on whether data synthesis, estimation of inestimable uncertainty, or data supplementation will be performed. The application of the created knowledge occurs when the knowledge gained is fed back into the drug development process to aid the understanding of the response surface of a drug under development. Communication is the key to the usage of the product of the knowledge creation process.

The rest of the paper is focused on the data supplementation approach for PK/PD knowledge creation.

DATA SUPPLEMENTATION

Data supplementation deals with the supplementation of data to enable the exploration of an aspect of the response

surface that may not have been targeted for exploration in a completed trial. It also deals with the supplementation of data in preclinical animal studies where the destructive nature of the sampling design does not permit the construction of individual profiles for inaccessible tissues. A motivating example that deals with a targeted aspect of an unexplored region of the response surface is discussed below to provide clarity on the approach. Data supplementation in the preclinical animal setting is beyond the scope of this paper.

Motivation for Data Supplementation

The motivation for data supplementation comes from the following:

- After data from a trial has been analyzed, it may become obvious that the dose range explored was limited, and more information (data) would be needed to gain an understanding of the effect of a dose or doses not studied.
- Abrupt cessation of a clinical trial could occur for nonclinical reasons, such as a nonclinical toxicology study finding. In such a situation, not all subjects would have completed the clinical trial—an incompletely observed study. The question arises as to what the responses of the subjects who could not complete the study would have been if the trial was not stopped abruptly. If a solution was found that provided an insight into what the study outcome would have been, the need for repeating such a study once the nonclinical problems are resolved could be obviated.
- Sometimes the clinical trial data do not lend themselves to the traditional PK/PD analyses. Consider a situation in which drugs are administered as combinations in a clinical trial due to their anticipated synergy, but the concentrations of the primary drug driving the effect is unavailable while that of the synergistic drug is available. In such a situation, a PK/PD model cannot be developed to characterize the interaction, but there is the need to characterize the effect that could be produced with a different dose of the interactor drug while the dose of the primary drug remains constant.
- A drug may be found, after a clinical trial, to appear to exhibit an inverted U-shaped response, and it is not clear whether a dose not studied in the trial could have produced an effect on the upswing of the dose-response curve that is more effective than a dose in the downswing of the dose-response curve.

Methodology for Data Supplementation

Multiple supplementation (MS) and its modification thereof—structure-based multiple supplementation (SBMS) approach—is proposed as a method for addressing the issue of data supplementation for the characterization of a targeted region (e.g., effect of a dose or dose range) of an unexplored response surface. The MS approach is an adaptation of the multiple imputation methodology used for augmenting data in missing data situations to enable data analysis on a complete data set. First the procedure for performing data supplementation is described, followed by a review of the multiple imputation (MI) methodology, and a description of the MS approach, and a discussion of SBMS in the context of a motivating example.

Data Supplementation Procedure

The procedure for data supplementation is as follows:

1. Statement of the objective of data supplementation for PK/PD knowledge creation;
2. Performance of PK knowledge discovery;
3. PK data synthesis for target dose groups;
4. Covariate data synthesis for virtual subjects in the target dose group(s);
5. Discovery of hidden knowledge from real data set to which supplemental data will be added;
6. Implementation of a data supplementation methodology (i.e., MS and its modification, SBMS); and
7. Discovery and communication of the created knowledge.

The MI methodology is first described to provide the framework for understanding the basis of it being adapted to create the MS methodology.

Multiple Imputation

Multiple imputation (MI), developed by Rubin (7,8), is a predictive approach to handling missing data in multivariate analysis. It blends both classic and Bayesian statistical techniques and relies on specific iterative algorithms to create several imputations. MI rectifies the major disadvantages (i.e., bias and lack of a measure of parameter estimation uncertainty) of single imputation methods (e.g., the use of a mean or median value or the last observation carried forward to fill in missing values) by replacing each missing value with a vector composed of $M \geq 2$ possible values (usually between 2 to 10, but commonly 5), to accurately reflect uncertainty and to preserve important data relationships and aspects of the data distribution. It requires that the analyst specifies an imputation model, imputes several data sets, analyzes them separately, and then combines the results. MI yields a single set of test statistics, parameter estimates, and standard errors.

The validity of the method hinges on how the imputations are generated. It is not possible to obtain valid inferences if imputations are created arbitrarily. On average, the imputation should give reasonable predictions for the missing data, and variability among them should reflect an appropriate degree of uncertainty. Rubin (9) provides technical conditions under which repeated-imputation method leads to frequency-valid answers. An imputation method that satisfies these conditions is said to be "proper" (9). Stated simply, procedures for imputation, whether based on explicit (parametric) or implicit (nonparametric) models, ignorable or nonignorable models, that incorporate appropriate variability among repetitions within a model are called "proper." "Ignorable missingness" occurs when the probability of a missing value is not dependent on the value itself, but may depend on the values of other variables in the data set (7). A variety of proper imputation methods based on both explicit and implicit models, including a fully normal model, the Bayesian bootstrap, and the approximate Bayesian bootstrap (ABB), have been studied by Rubin (10). An imputation model must preserve all important associations among variables in the

data set, including interactions if they will be part of the final analysis. Also, the dependent variable must be included in the model (11).

The MI Paradigm: Parametric Bayesian Models

Rubin's suggested Bayesian approach to MI was popularized by Schafer (11), who provided detailed algorithms for creating MIs in different situations. Suppose, in general, that $Y = (y_1,y_2,\ldots,y_n)$, where the first a values $[Y_{obs} = (y_1,y_2,\ldots,y_a)]$ are actual observed values and the remaining values $[Y_{mis} = (y_{a+1},y_{a+2},\ldots,y_n)$ are missing at random. $Y = (Y_{obs},Y_{mis})$ follows a parametric model $Y \sim P(Y|\theta)$, where θ is the unknown parameter, or a vector of parameters in the multivariate case, that we are ultimately interested in (e.g., mean, variance, or shape that describes the response surface). θ is assumed to have a prior distribution and Y_{mis} is ignorably missing. MIs are Bayesianly proper if they are independent realizations of $P(Y_{mis} \mid Y_{obs})$, the posterior predictive distribution of the missing data under some complete data model and prior. $P(Y_{mis} \mid Y_{obs})$ may be written as

$$P(Y_{mis} \mid Y_{obs}) = \int P(Y_{mis} \mid Y_{obs},\theta)P(\theta \mid Y_{obs})d\theta$$

the conditional predictive distribution of Y_{mis} given θ, averaged over the observed-data posterior of θ. Thus, Bayesianly proper imputations reflect uncertainty about Y_{mis} given the parameters of the complete-data model, as well as uncertainty about the unknown model parameters. The resulting MIs are appropriate under an assumption of ignorability because $P(Y_{mis} \mid Y_{obs})$ does not rely on the pattern of the observed response. Thus, an imputation for Y_{mis} can be described in two steps: first by simulating a random draw of the posterior distribution of the unknown parameter $\theta^* \sim P(\theta \mid Y_{obs})$ and followed by a random draw of the missing values from their conditional predictive distribution $Y^*_{mis} \sim P(Y_{mis} \mid Y_{obs},\theta^*)$.

For some cases, the posterior distribution of θ is not straightforward, due to nonstandard distribution that can not be easily simulated. Rubin (9) introduced a few general strategies for approximating draws for the posterior distribution of θ, including large-sample normal approximations and importance resampling.

The parametric Bayesian approach could be adopted for data supplementation. However, we choose to adapt the nonparametric approximate Bayesian bootstrap approach, which makes minimal distributional assumptions for data supplementation, and this is discussed next.

Nonparametric Approximate Bayesian Data Supplementation Method

Rubin (9) described a simple method for MI called the approximate Bayesian bootstrap (ABB). This approach makes it possible to generate proper imputation for Y_{mis} with minimal distributional assumptions. We have adapted this approach for MS, making it possible to generate "proper" supplementation for Y_{supp} with minimal distributional assumptions. To illustrate the ABB approach for MS, consider a collection of n units with the same value of covariates $\boldsymbol{X}$, where a subjects were observed and $n_{supp} = n - a$ subjects (virtual) with values to be supplemented. The ABB creates M ignorable repeated supplementations from $m = 1, \ldots M$ as follows: a) create a new pool of Y^*_{obs} by sampling a values from $Y_{obs} = (y_1, y_2,\ldots,y_a)$ with replacement, and b) select a set of n_{supp} possible values from Y^*_{obs}, again with replacement. By drawing n_{supp} supplemented values from a *possible* sample of Y^*_{obs} values rather than from the Y_{obs} values, the ABB approach generates appropriate between-supplementation variability, at least assuming large sample random samples given covariates $\boldsymbol{X}$. This is akin to the generation of imputation variability, assuming large sample random samples as demonstrated by Rubin and Schenker (12).

Combining Estimates

Following the approach used in the MI paradigm, after M supplementations have been created for a data set, they are then analyzed using standard PK/PD or statistical package. There are now M completed data sets containing the observed values and the supplemented values instead of one. The PK/PD or statistical analysis must be done M times, once on each complete data set. Across M data sets the results will vary, reflecting the uncertainty due to supplemental observations. The M complete data analyses are combined to create one repeated-supplementation inference.

Let $\hat{\Theta}_m$ and U_m, $m = 1, \ldots M$, be M complete supplemented data estimates and their associated variances for a parameter Θ, calculated from the M data sets completed by repeated supplementations under one model. For instance, $\Theta = \beta$, $\hat{\Theta}_m$ is the least squares estimate of β, and U_m is the weighted residual mean square error. The repeated supplementation estimate of Θ is the mean of the complete data estimates:

$$\overline{\Theta} = \sum_{m=1}^{M} \hat{\Theta}_m \Big/ M$$

There are two components of the variability associated with this estimate:

The average within-supplementation variance,

$$\overline{U} = \sum_{m=1}^{M} U_m \Big/ M$$

and between-supplementation component,

$$B = \sum_{m=1}^{M} (\hat{\Theta}_m - \overline{\Theta})^2 \Big/ (M-1)$$

The total variability associated with $\overline{\Theta}$ is given by

$$T = \overline{U} + (1 + M^{-1})\, B$$

Inference can be made using Θ, T, and a distributional assumption. For example, if Θ is a scalar quantity, the approximate reference distribution for interval estimates and significance tests is a t distribution:

$$(\Theta - \overline{\Theta})\, T^{-1/2} \sim t_\nu$$

where the degrees of freedom, ν, are given by

$$\nu = (M-1)\,(1 + r^{-1})^2$$

with $r = (1 + M^{-1})B/\overline{U}$ *(12).*

Thus, a $100(1 - \alpha)\%$ interval estimate for $\overline{\Theta}$ is

$$\overline{\Theta} \pm t_{v,1-\alpha/2} \sqrt{T}.$$

The between and within subject ratio, r, estimates the population quantity $\gamma/(1 - \gamma)$, where γ is the fraction of information about Θ supplemented.

STRUCTURE-BASED MULTIPLE SUPPLEMENTATION: A MOTIVATING EXAMPLE

This example illustrates how knowledge can be created using a combination of data synthesis, structure revelation, and multiple supplementation (MS) techniques. Because data supplementation was performed based on the structure revealed from the data, as discussed later, this modification of the MS approach is termed *structure-based multiple supplementation* (SBMS) approach. A parallel dose efficacy study of a drug in development was performed with three dose levels: placebo, 200 mg, and 600 mg. Subjects were sampled for population pharmacokinetic and efficacy analysis. The objective of this PK/PD knowledge creation investigation was to determine a likely treatment outcome if subjects were randomized to a 100-mg dose group that was not studied in an already completed trial. The 100-mg dose group is hereafter referred to as the target dose group.

Prior to performing PK/PD knowledge creation, PK knowledge discovery (6) was performed. Thirty-five subjects were administered the test drug on a three times daily basis for 28 days. Eighteen and seventeen subjects were randomized to receive 200 and 600 mg of test drug orally. These subjects provided 974 concentrations, yielding an average of 27.8 concentrations/subject over a 28-day period. The mean (SD) age, weight, and height of subjects were 45 (6.3) years, 85 (31.3) kg, and 171.6 (12.3) cm, respectively. There were 21 male and 13 female subjects, 25 whites, 5 blacks, and 4 Hispanics.

The population PK model was developed as a consequence of PK knowledge discovery performed on the data described above. Briefly, graphical displays were used for structure revelation and hidden patterns in the data. Thereafter, one- and two-compartment pharmacokinetic models with first-order input were tested for their ability to appropriately characterize the PK of the drug using the NONMEM software. The PK data were best described with the two-compartment model incorporating a first-order input. The parameters of the model were absorption rate constant, apparent volume of the central compartment, apparent volume of the peripheral compartment, intercompartmental clearance, and apparent clearance. Empirical individual Bayesian *post hoc* parameter estimates were obtained and subjected to more exploratory data analysis [i.e., graphical analysis and generalized additive modeling (GAM)] for initial covariate selection. The GAM analysis was coupled with bootstrap replication stability (13) to select covariates with inclusion frequency of ≥50% from 100 nonparametric bootstrap replicates. The covariates—age, and dose level—selected by GAM were used to create a full model in NONMEM from which an irreducible final model was obtained by backward elimination. The irreducible model for the key parameter of interest, apparent clearance, included age and dose level as significant predictors. Thus,

$$CL/F = 30.2 - 3.22 \times (\text{Age-Median}) \times IND + 0.14 \times DL$$

where IND takes on the value of 0 if Age is greater than the median age and 1 otherwise, and DL is dose level.

The model was used to simulate concentrations from which area under the plasma concentration-time curve (AUC) was computed and compared with AUC calculated from model-based parameters as a means of checking the posterior predictive performance of the model. Figure 1 shows the distribution of AUC values for observed (model-based) and simulated data for the two studied doses and the 100-mg target dose. The results of 10 replications out of 100 are shown in the figure for illustrative purposes. The median simulated AUC for the 600-mg dose was similar to the model calculated AUC, and more than 88% of the AUC values from the simulated data overlap the model-calculated AUC. The median AUC values obtained from the simulated data for the 200-mg dose were slightly biased (18%) with 68% of the values overlapping the model-calculated AUC values. The Kolmogorov-Smirnov goodness-of-fit test, a two-sample comparison test, was used to perform the posterior predictive check for each replicate. The null hypothesis was that the observed (model-based) AUC distribution and the simulated ones were equal, with the alternative that they were not. The ranges of p value obtained across replicates for the 200- and 600-mg doses were 0.11 to 0.75 and 0.24 to 1, respectively, indicating similarity in the distributions. Although the model performed better in predicting AUC with the 600-mg dose than with the 200-mg dose, the population PK model developed did provide a reasonably adequate description of the data and was later used to simulate PK profiles for virtual subjects used in data supplementation described below.

With PK knowledge discovery performed, the PK/PD knowledge creation was then performed in two phases as follows:

Phase I consisted of 3 steps:

1. Simulation of subjects with demographics similar to those in real study data set. Briefly, data synthesis of covariates for the virtual study was done through a resampling with replacement approach to ensure that the covariate distributions in each virtual study were similar to the real study. The

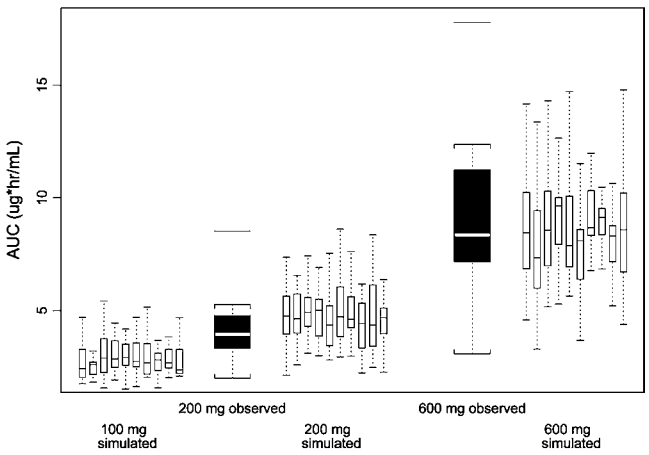

Fig. 1. Distributions of AUC values for 200- and 600-mg dose groups studied with parallel comparisons of those obtained via posterior predictive performance check. AUC values from PK data simulated for the target dose of 100 mg are also included for comparison. "Observed" in the figure refers to population PK model-based AUC.

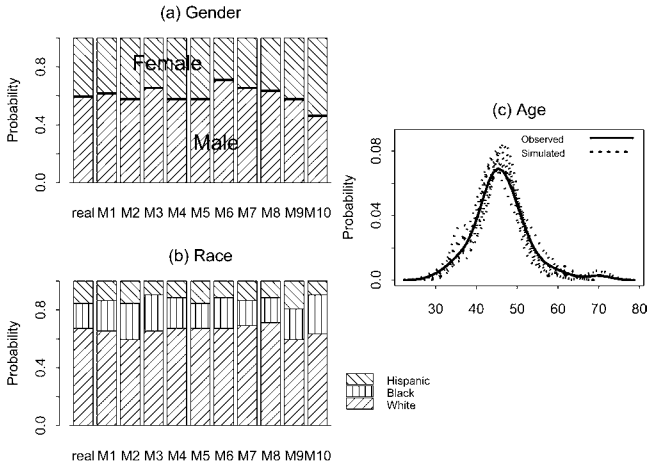

Fig. 2. Comparison of distributions of demographic variables between the real and simulated data sets. M1 to M10 represent the number of replications used for data supplementation.

demographic data from the real study was examined for correlation between covariates. There were no significant correlations between age and gender, and age and race. Also, gender and race were not correlated. However, age and weight were correlated. Given the total number of subjects, n, in the data set, a covariate vector such as gender was resampled with replacement from the observed data so that the proportion of males and females in the simulated data set was similar to that in the real data. This procedure was repeated for the other uncorrelated covariates. Where the covariates were correlated, the resampling was done at the subject level to maintain the correlation structure. This resampling with replacement approach ensures equal probabilities for each element (covariate) of the population. The above algorithm was replicated for each virtual study used in the multiple data supplementation step discussed latter. Figure 2 illustrates the distributions of some of the covariates. It is worth noting the similarity between the distributions of the resampled covariates and those from the real study.

2. Simulation of pharmacokinetic profiles for subjects from step 1 using a population pharmacokinetic model developed from data obtained from previously completed trial and computation of exposure metrics and PK parameters for the simulated subjects. The distribution of AUC values, for instance, the target dose of 100 mg, is shown in Fig. 1. These AUC values obtained from the first 10 replications were used for data supplementation.

3. Combination of individual PK (exposure) variables from virtual subjects with subject specific covariate data together with the real data set (including the PD-biomarker response data) to create a PK/PD knowledge creation data set.

Phase II was performed in 2 major steps:

1. Performance of data structure analysis on real study data (untransformed and transformed) to reveal hidden structure, patterns, and relationships in the data set. This involves data visualization (graphing and fitting) and exploratory modeling (e.g., tree-based modeling).

2. MS is used to generate ($M = 10$) baseline biomarker values for simulated subjects in the target dose group in which knowledge is to be created. MS is performed based on the revealed data structure. Figure 3 provides the schema of an example of the SBMS approach. In the example under consideration, the target dose group was partitioned into two groups: group A: subjects with higher biomarker baseline values and younger age (likely responder group); and group B: the remainder of the subjects (the likely nonresponder group). Reduction of biomarker levels from baseline value was an indication of subject response to therapy. Basically, the data supplementation for group A proceeded as follows: a) the slope from day 0 to day 8 was supplemented from the real data that contained subjects matching the subpopulation criteria from day 0 to day 8; and b) slopes for other time periods (i.e., day 8 to 15 and day 15 to day 28) were supplemented from all available data at the same time periods.

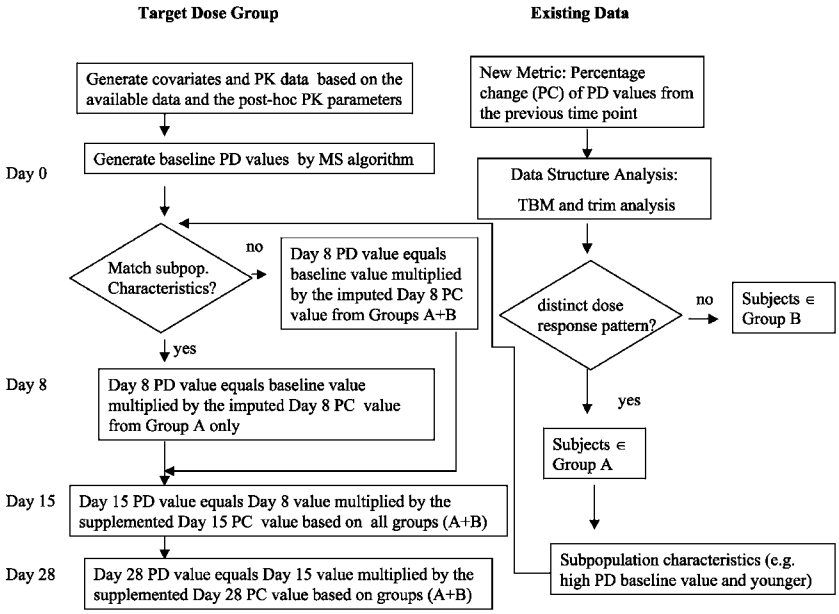

Fig. 3. A schematic of the structure-based multiple supplementation (SBMS) approach.

For group B subjects, biomarker responses were supplemented from all available PD data to reflect the overall uncertainty. Altogether, 10 replicates of target group data sets were created for the target dose group.

Figure 4 shows the transformed data distributions of percentage changes from baseline values (i.e., slope) across replicates for the 100-mg target dose group from day 0 to day 8. In Fig. 4a, all subjects are included, but Fig. 4b contains the responder subpopulation only. It can be observed that subjects who met the responder criteria (Fig. 4b) had steeper slope values; the majority were in the –0.5 to –1 range.

After the creation of the biomarker data, each of 10 replicate data sets for the target 100-mg dose group was subjected to modeling and the results combined for PK/PD knowledge creation on the performance of the 100-mg dose level. The details of the modeling and results thereof are beyond the scope of this chapter. However, Fig. 5 presents the data created for the target dose group in addition to the real data that were collected for the other those groups that were studied. The results were consistent with the pharmacology of the drug. The supplemented biomarker response for the 100-mg target dose group appeared better than that observed for the 600-mg dose in the responder population as revealed by the slope, but comparable with the 200-mg dose (Fig. 5). The knowledge created about the performance of the target 100-mg dose was communicated to the development team for the design of a future trial.

DISCUSSION

The success rate of new chemical entities (NCEs) is anything but stellar (14). In 1987, the cost of bringing a new drug into the market was $237 million as opposed to $802 million in 2000 (15). By the end of 1999, 21% of the NCEs with investigational new drug applications (INDs) filed from 1981 to 1992 had been approved for marketing in the United States (16). Of those that failed in the period from 1987 to 1992, 38% of the NCEs failed because of efficacy (activity too weak or lack of efficacy), 34% on economics (commercial market too limited or insufficient return on investment), 20% because of safety (human or animal toxicity), and the rest for nonspecific reasons (16). What is becoming increasingly clear is that traditional drug development approaches are unlikely to succeed in the future given the economics of drug development—a low probability of success coupled with increasing product development times means decreased sales time after market launch and lower return on investment for pharmaceutical companies.

To speed drug development, sophisticated new technolo-

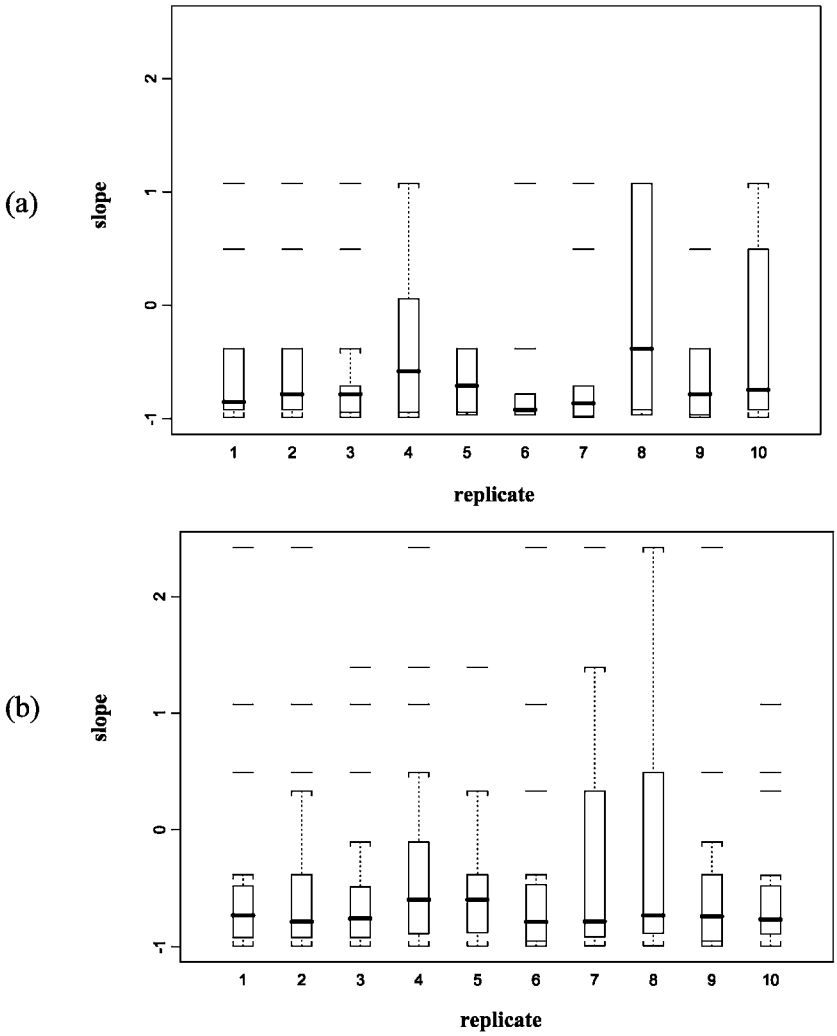

Fig. 4. Day 0 to 8 slope distributions across multiple supplementation replicates for the target dose level: (a) all subjects, and (b) subjects in the responder subgroup.

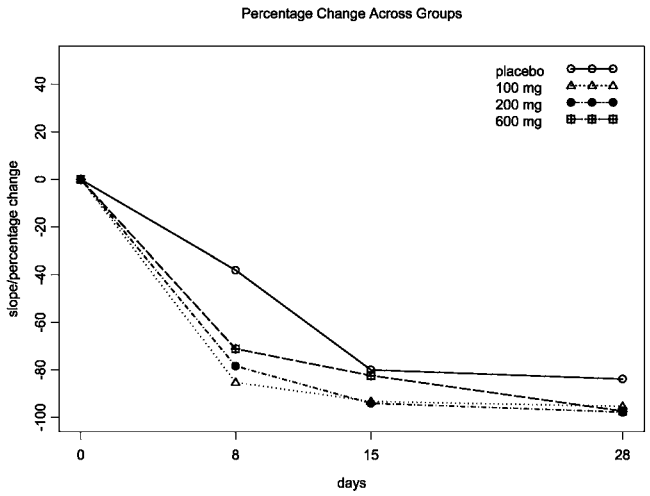

Fig. 5. Slope distributions across days among different dose groups for subjects in the responder subgroup. Note that 100 mg is the targeted dose group.

gies and approaches in the discovery and design of new drugs are replacing the traditional methods of discovery. Increasingly, however, a pharmacometrically guided approach is being applied to drug development. The need to get the most knowledge from every drug development study that is performed cannot be overemphasized in this day and age of spiraling drug development cost. This need has lead to the development of PK/PD knowledge creation in general, and data supplementation in particular. In the application example, the nature of the response to a targeted dose was obtained. The drug effect that would have been produced if the target dose of 100 mg was studied would have been better than that produced by the 600-mg dose but comparable with that produced by the 200 mg in the responder group. Thus, the drug was postulated to have an inverted U-shaped dose-response curve with doses above 200 mg producing lesser effect than the 200-mg dose.

The application of PK/PD knowledge creation and implicitly knowledge discovery during drug development will optimize the drug development process and promote rational pharmacotherapy. Concerning drug development, Minto and Schnider (17) have stated, "Rapidly evolving changes in health care economics and consumer expectation make it unlikely that traditional drug development approaches will succeed in the future. A shift away from the narrow focus of rejecting the null hypothesis toward a broader focus of seeking to understand the factors that influence the dose-response relationship together with the development of the next generation of software based on population models should permit a more efficient and rational drug development programme." The drug development process can be improved by implementing knowledge-driven development strategies founded on powerful, informative, and robust clinical trials. PK/PD knowledge creation and its companion knowledge discovery processes play pivotal roles in the generation and extension of knowledge and therefore can be influential in bringing efficiencies to the drug development process.

PK/PD knowledge creation via data supplementation results in the further acquisition of knowledge beyond that embedded in one's data. When data supplementation for PK/PD knowledge creation is implemented, the result is a greatly improved understanding of the response surface because of the knowledge created. This in turn leads to the efficient design studies.

The approach to data supplementation described in this paper can only be implemented if the conditions stipulated under the section "Motivation for Data Supplementation" are met. The MS approach and its modification, SBMS, are an adaptation of the MI paradigm with the focus being data supplementation and not "missingness." From Figs. 4a and 4b, it is obvious that at least five replications are sufficient for obtaining robust data from the data supplementation process. MS should not be confused with missingness and the conditions that must be satisfied for before performing MS. With MS, the focus is on characterizing an unexplored region of the response surface that was not the subject of focus in an already completed study. With the MS approach, data are generated for knowledge creation or acquisition, whereas MI deals with filling in missing data to create a complete data set for analysis.

It must be stated that in the current climate, insufficient attention is given to knowledge-based drug development. The process of drug development can be no better than the knowledge on which it is based. Without adequate knowledge, it is impossible to have a thorough understanding of one's drug with the consequent compromising of the optimal development strategy. PK/PD knowledge creation through data supplementation, with knowledge discovery being an implicit component of it, results in gaining knowledge about a targeted region of a response surface that was not previously studied in a completed study—this gain without expending resources in conducting a new study. This provides the drug developer with the wisdom to make the right decisions about future trials and the strategic path for the development of a drug.

REFERENCES

1. P. J. Williams, C. J. Godfrey, A. Roy, H.-M. Chu, and E. I. Ette. Pharmacokinetics/pharmacodynamics knowledge discovery and creation during drug development. In P. Bonate and D. Howard (eds.), *Industrial Pharmacokinetics.* AAPS Press, 2004. pp. 163–202.
2. D. R. Jones. Computer simulation as a tool for clinical trial design. *Int J Bio-Med Comp.* **10**:145–150 (1979).
3. K. L. Lee, M. Frederick, C. F. Starmer, P. J. Harris, and R. A. Rosati. Clinical judgment and statistics: lessons from a simulated randomized trial in coronary artery disease. *Circulation* **61**:508–515 (1980).
4. E. I. Ette and L. C. Onyiah. Estimating inestimable standard errors: the bootstrap and Winsorization. *Eur. J. Drug Metab. Pharmacokinet.* **27**:213–224 (2002).
5. L. B. Sheiner. Learning versus confirming in clinical drug development. *Clin. Pharmacol. Ther.* **61**:275–291 (1997).
6. E. I. Ette EI, P. Williams, H. Sun, E. Fadiran, F. O. Ajayi, and L. C. Onyiah. The process of knowledge discovery from large pharmacokinetic data sets. *J. Clin. Pharmacol.* **41**:25–34 (2001).
7. D. B. Rubin. Inference and missing data. *Biometrika* **63**:581–582 (1976).
8. D. B. Rubin. Multiple imputation after 18+ years. *J. Am. Stat. Assoc.* **91**:473–489 (1996).
9. D. B. Rubin. *Multiple Imputation for Nonresponse Surveys.* Wiley, New York, 1987.
10. D. B. Rubin. The Bayesian bootstrap. *Ann. Stat.* **9**:130–134 (1981).
11. J. L. Schafer. Analysis of incomplete multivariate data. Chapman & Hall/CRC, Boca Raton, FL, 2000.
12. D. B. Rubin and N. Schenker. Multiple imputation for inter-

val estimation from simple random samples with ignorable nonresponse. *J. Am. Stat. Assoc.* **81**:366–374 (1986).
13. E. I. Ette. On population model stability and performance. *J. Clin. Pharmacol.* **37**:486–495 (1997).
14. M. I. Kleinberg and L. A. Wanke. New approaches and technologies in drug design and discovery. *Am. J. Health Syst. Pharm.* **52**:1323–1336; 1341–1343 (1995).
15. C. Connolly. Price tag for a new drug: $802 million; findings of Tufts University study are disputed by several watchdog groups. *Washington Post*, 1 December 2001.
16. J. A. DiMasi. Risks in new drug development: approval success rates for investigational drugs. *Clin. Pharmacol. Ther.* **69**:297–307 (2001).
17. C. Minto and T. Schnider. Expanding clinical applications of population pharmacodynamic modeling. *Br. J. Clin. Pharmacol.* **46**:321–333 (1998).