Journal of Pharmacokinetics and Pharmacodynamics, Vol. 33, No. 6, December 2006 (© 2006)
DOI: 10.1007/s10928-006-9033-1

Simulation of Correlated Continuous and Categorical Variables using a Single Multivariate Distribution

Stacey J. Tannenbaum,[1,5] Nicholas H. G. Holford,[2,3] Howard Lee,[2] Carl C. Peck,[2] and Diane R. Mould[4]

Received 25 March 2006—Final 22 August 2006—Published online October 12, 2006

Clinical trial simulations make use of input/output models with covariate effects; the virtual patient population generated for the simulation should therefore display physiologically reasonable covariate distributions. Covariate distribution modeling is one method used to create sets of covariate values (vectors) that characterize individual virtual patients, which should be representative of real subjects participating in clinical trials. Covariates can be continuous (e.g., body weight, age) or categorical (e.g., sex, race). A modeling method commonly used for incorporating both continuous and categorical covariates, the Discrete method, requires the patient population to be divided into subgroups for each unique combination of categorical covariates, with separate multivariate functions for the continuous covariates in each subset. However, when there are multiple categorical covariates this approach can result in subgroups with very few representative patients, and thus, insufficient data to build a model that characterizes these patient groups. To resolve this limitation, an application of a statistical methodology (Continuous method) was conceived to enable sampling of complete covariate vectors, including both continuous and categorical covariates, from a single multivariate function. The Discrete and Continuous methods were compared using both simulated and real data with respect to their ability to generate virtual patient distributions that match a target population. The simulated data sets consisted of one categorical and two correlated continuous covariates. The proportion of patients in each subgroup, correlation between the continuous covariates, and ratio of the means of the continuous covariates in the subgroups were varied.

[1]Novartis Pharmaceuticals Corp., One Health Plaza 435/1125, East Hanover, NJ 07936, USA.
[2]Center for Drug Development Science, UC Washington Center, School of Pharmacy, University of California San Francisco, 1608 Rhode Islands Road, NW, Washington, DC 20036, USA.
[3]Department of Pharmacology and Clinical Pharmacology, University of Auckland, 85 Park Rd, Private Bag 92019, Auckland, New Zealand.
[4]Projections Research, Inc., 535 Springview Lane, Phoenixville, PA 19460, USA.
[5]To whom correspondence should be addresed. E-mail: stacey.tannenbaum@novartis.com

During evaluation, both methods accurately generated the summary statistics and proper proportions of the target population. In general, the Continuous method performed as well as the Discrete method, except when the subgroups, defined by categorical value, had markedly different continuous covariate means, for which, in the authors' experience, there are few clinically relevant examples. The Continuous method allows analysis of the full population instead of multiple subgroups, reducing the number of analyses that must be performed, and thereby increasing efficiency. More importantly, analyzing a larger pool of data increases the precision of the covariance estimates of the covariates, thus improving the accuracy of the description of the covariate distribution in the simulated population.

KEY WORDS: covariate; covariate distribution modeling; continuous covariate; categorical covariate; multivariate distribution; clinical trial simulation.

INTRODUCTION

Clinical trial simulation (CTS) can be a valuable tool to improve drug development (1–3). By synthesizing the available knowledge about the drug, patients, and clinical program (e.g., pharmacokinetics and pharmacodynamics, disease progress, demographics) into a stochastic model, the user can investigate, in silico, aspects of the clinical study plan (dosing regimens, study designs, patient populations, formulations), allowing the clinical team to make rational, informed decisions with regards to optimizing the development plan of a new compound (4–8).

A clinical trial simulation model consists of three main components (1): a clinical trial execution model, an input–output (IO) model, and a covariate distribution model. The execution model describes aspects of the study conduct such as compliance with dosing schedules, and subject dropouts. The IO model is a collection of models describing the disease progress during the study period, and the pharmacokinetics and pharmacodynamics of the drugs being tested. The covariate distribution model incorporates patient-specific factors that may account for inter-individual differences in observed pharmacokinetics and pharmacodynamics and contribute to variability in individual parameter values. Based on the established or hypothesized impact of the covariates on the IO model, the simulated covariate information is then used to predict IO model parameters for a virtual patient with a particular combination of demographics and characteristics.

Covariate distribution modeling can be used to generate virtual patients for clinical trial simulation (3,9). Each patient is represented by a set of intrinsic or extrinsic factors (called a covariate vector) which collectively describe the characteristics of the patient. Useful covariates typically include demographics (age, weight, sex, race), concomitant drug use (which may also include abused drugs, tobacco and alcohol), and disease risk or health status biomarkers (e.g., blood pressure, cholesterol concentrations, creatinine clearance, liver enzymes, disease severity). Note that patient covariates may be continuous

(such as age and weight) or categorical (such as sex, race, or smoking status). These covariates are frequently correlated between individuals (e.g., women are more likely to weigh less and have lower creatinine clearances than men). Since the covariates are used to predict elements of the IO model that influence a patient's trial outcome, it is critical that the covariates associated with each virtual patient be realistic and consistent with the projected patient population. Therefore, some care should be given to the development of the covariate distribution model.

There are a number of techniques that use the covariates of an existing patient population (e.g., a patient population with the same indication, or patient information from a previous study of the same drug) to create new virtual patient populations for clinical trial simulation (9). The simplest method is to sample complete patient covariate vectors from observed values in the existing database (also called the empirical distribution), with or without replacement of that vector in subsequent sampling. The benefit of sampling from an empirical distribution is that covariate combinations are guaranteed to be realistic, as they are extracted directly from real patient data. However, no new patient covariate vectors can be created using this approach.

Rather than sampling complete vectors from single subjects, the individual empirical distributions of each covariate can be used to create vectors that do not exist in the empirical database. In this method, covariates are sequentially sampled from their individual empirical distributions, with each subsequent covariate chosen from a constrained set of values based upon previously selected covariates (e.g., after randomly sampling age from the observed age distribution, one would then randomly sample creatinine clearance from its observed distribution; however, the choice of values would be limited to creatinine clearance measurements obtained from patients with the age chosen in the first step). Such "conditional distributions" will preserve the correlation between the covariates. It should be noted, however, that as each additional covariate is selected, subsequent distributions become increasingly constrained, potentially limiting values to a highly restricted subset; shuffling the selection order may partly alleviate this problem (9). In addition, the process of sequentially selecting covariates can be computationally inefficient.

Random sampling of covariate vectors from a multivariate normal distribution (MVND, Fig. 1) preserves the benefits of the previously described covariate profile generation methods (generation of unique subjects with realistic covariate vectors), while reducing their limitations (e.g., computational inefficiency and sampling from overly constrained distributions). A MVND is represented by two parameters: a vector of means of the individual covariates, and matrix consisting of the variances of the

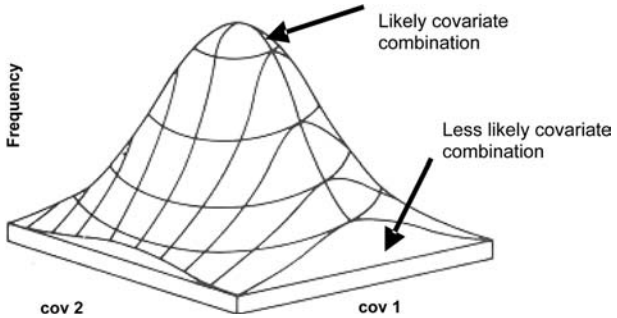

Fig. 1. Illustration of multivariate normal distribution for two covariates (cov1 and cov2). Covariate combinations that occur naturally in the target population have a high probability of being selected by the covariate distribution model, whereas unrealistic or physiologically impossible combinations are selected with lower frequency.

covariates along the main diagonal, and the covariances between each pair of covariates in the other matrix positions (10).

Two important assumptions fixed in the definition of an MVND must be considered. First, all covariates in the MVND are assumed to follow the same known distribution (e.g., normal or log-normal). Second, while covariance defines the basic association between two covariates, it is not sufficient to fully define the *shape* of the relationship; sampling from a MVND will always result in linearly related simulated covariates. Thus, regardless of the observed distributions of the covariates, and the shapes of the relationships between them in the empirical distribution, covariates sampled from an MVND based upon this data will be normally (or log-normally) distributed and linearly related; the simulated results may therefore only approximate the original target population. It should be noted, however, that most common covariates, such as age and weight, are generally normally or log-normally distributed; in addition, within normal ranges of covariate values, it is usual to see a linear relationship between common covariates. Therefore, because the MVND defines the individual covariate distributions as well as maintains the systematic relationship between the covariates, the generated covariate vectors should be physiologically realistic.

Software packages such as NONMEM (11) and Pharsight® Trial Simulator (Version 2.1.2, Pharsight Inc, Mountain View, CA) allow samples to be obtained from multivariate normal distributions (MVND), but this can only be accomplished when all covariates are continuous. Because categorical covariates are not continuous they have not previously been considered for inclusion in MVNDs. However, the method for covariate distribution modeling that will be introduced creates a single MVND

which includes both continuous and categorical covariates. This is accomplished by deliberately ignoring the categorical nature of a covariate and treating it as if it arose from a continuous distribution. Evaluation of the properties of this method (the Continuous method) was undertaken by comparing it with a standard method (the Discrete method) with respect to their abilities to generate virtual patient distributions that match a real or simulated target population.

METHODS

Discrete Method

A commonly used method for dealing with both continuous and categorical covariates is to use a separate MVND for each unique combination of categorical covariate values. This will henceforth be designated the *Discrete* method. For example, if sex and smoking are the two categorical covariates, the population is divided into four groups (female smokers, female nonsmokers, male smokers, and male nonsmokers). Subgroup specific MVNDs are then created from which continuous covariates (e.g. age, weight) are sampled. Although the Discrete method is frequently used, there are significant limitations, which arise from subdividing the patient population.

First, the Discrete method may be impractical to implement when there are multiple categorical covariates (e.g. sex $\times 2$, smoking $\times 3$, race $\times 4$, disease status $\times 5$ would lead to 120 separate MVNDs). Even if it were feasible to simulate this many MVNDs, estimation of their parameters may be impossible because of limited empirical observations of the continuous covariates in each of the categorical subgroups. When there are too few patients in a subgroup, there may be insufficient data to create a reliable MVND; specifically, if there less than $N + 1$ subjects in a subgroup (where N is the number of covariates in the MVND), the variance–covariance matrix that is generated will be singular. A worst case scenario is a subgroup in which there are no patients in the empirical distribution, yet patients with this combination of categorical covariates could potentially be enrolled in a future clinical trial. Because the relationship between the continuous covariates in this subgroup is unknown, it is impossible to determine if the simulated patient covariate vectors are appropriate for this patient group.

Although there may be no data about the association between covariates (continuous or categorical) for a specific subgroup, it seems reasonable to assume that the variance structure may be similar to those

observed in subgroups occurring more frequently in the empirical distribution. Thus, we propose an approximate method to capture these correlations in order to simulate correlated categorical and continuous covariates. The method can be applied using estimates from a sparse (for some combinations) empirical distribution.

Continuous Method

In the *Continuous* method the parameters of a single MVND are estimated by treating all categorical covariates as if they are continuous values, a procedure seen commonly in statistical simulation (12–15). In order to constrain all biological covariates to be positive, we typically assume a log-normal multivariate distribution. Thus, the MVND variance–covariance matrix is defined in terms of the logarithms of the covariate values. Likewise, categorical values must all be coded to possess positive values. Complete patient covariate vectors (both continuous and categorical covariates) are then sampled from a single MVND; because the sampled values are logarithmic, each component of the vector is then exponentiated to obtain the true covariate values. Note that because categorical covariates are sampled from a continuous MVND, sampled values for the virtual patients for these covariates will therefore be nondiscrete (e.g., if nonsmokers = 1 and smokers = 2 in the empirical distribution, a value such as 1.3 would be a possible value for smoking status in vectors sampled from a MVND). These continuous values must then be mapped to discrete categorical values, based on a continuous critical value (CrV).

The CrV is determined from the inverse of the lognormal cumulative distribution with a given mean, standard deviation, and cumulative probability (16), according to the following equation:

$$\mathrm{CrV}(\mu, \sigma, P_i) = e^{\mu+\sigma\cdot\mathrm{NORMINV}(P_i)}. \quad (1)$$

For a categorical covariate X with k discrete values, $\mu = \mathrm{mean}(\ln(X))$, $\sigma = \mathrm{SD}(\ln(X))$, P_i is the proportion of subjects in the empirical distribution with categorical value $X_i (i \leq k)$, and NORMINV is the inverse of the standard normal distribution. Figure 2 illustrates the CrV calculation for an example of a categorical covariate, smoking status, with three levels. The data set contained 18.6% nonsmokers, 49.3% former smokers, and 32.1% smokers, coded 1, 2, and 3, respectively. The corresponding cumulative probabilities (P_i) were 0.186, 0.679 (0.186 + 0.493), and 1. The original smoking status values were log-transformed and the mean and standard deviation of the transformed data were determined ($\mu = 0.694$, $\sigma = 0.378$). These values were used (Eq. 1) to determine two CrV values (1.43 and 2.39). Thus, if the continuous value for smoking status was

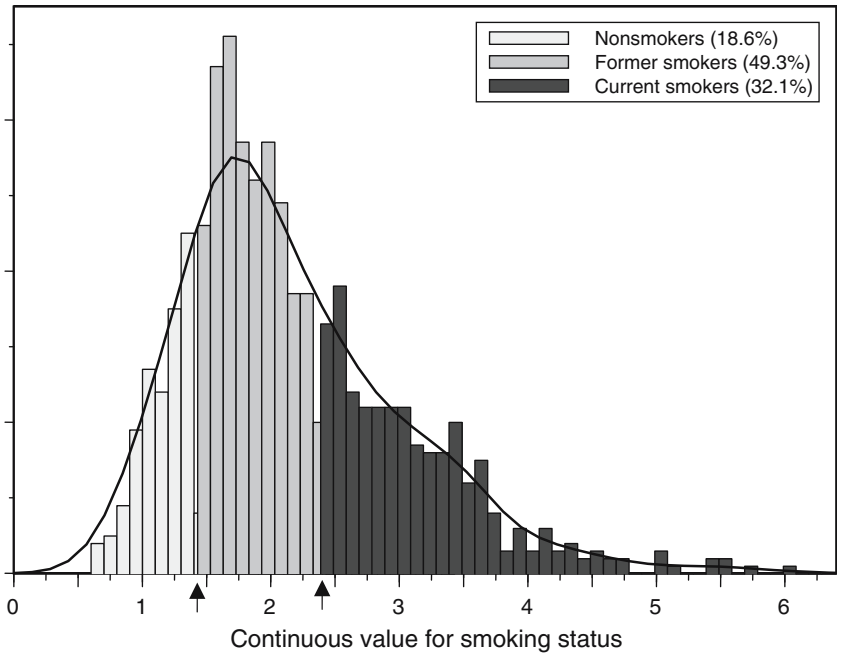

Fig. 2. Illustration of the calculation of discrete values for smoking status with three possible values (nonsmoker, former smoker, current smoker). The mean and standard deviation define the log-normal probability distribution of smoking status in the empirical distribution, as if it was a continuous covariate. The histogram is derived from an empirical distribution used to estimate the continuous distribution parameters. The cumulative probability (P) corresponds to the area under the probability distribution curve (solid line). The CrVs are indicated by arrows on the x-axis.

less than or equal to 1.43, the subject would be designated a nonsmoker; a value between the two CrVs would designate a former smoker, and if greater than 2.39, the subject would be designated a smoker.

Qualification of Methods

The performances of the Discrete and Continuous methods were evaluated based upon their abilities to reproduce the summary statistics of target population covariate distributions, using both real and simulated populations. All simulations were performed using Trial Simulator, and statistics were determined using SPLUS® (Version 6.0, Insightful Corp., Seattle, WA).

For the Continuous method, the parameters of a single MVND were estimated using all covariates (continuous and categorical) from the real

or simulated data. The following steps were carried out in S-PLUS. First, summary statistics (geometric mean, minimum, and maximum) were computed for each covariate. All covariates were log-transformed (for categorical covariates with a value of zero in the empirical distribution, 1 was added to all values prior to log-transformation). The variance–covariance matrix of the transformed values was then determined. The summary statistics and variance–covariance matrix were entered into Trial Simulator to define the MVND. All covariates were classified as continuous, regardless of their original type (categorical, continuous) in the empirical distribution. After the covariate vectors were sampled from the MVND, each element was exponentiated. The "continuous" categorical covariate values were then discretized based upon the appropriate CrV.

For the Discrete Method, the values from the real or simulated data were subset into groups corresponding to each unique combination of categorical covariates. In Trial Simulator, each covariate was classified as categorical or continuous according to its type in the empirical distribution. The method outlined in the previous paragraph was then applied for the continuous covariates in each subgroup.

1000 subjects were simulated using both the Discrete and Continuous methods. The population summary statistics and distributions of continuous covariates and proportions of categorical covariate values generated by both methods were compared to those of the corresponding "observed" (real or simulated) data. In addition, a test of the method's ability to preserve the correlation coefficients between the covariates was examined. A variance–covariance matrix was created from the simulated population data, and compared to the variance–covariance matrix obtained from the original data; the percent difference between the values in the same positions in the two matrices was calculated.

For each original data set (real, simulated), the Discrete and Continuous methods were replicated 10 times. With 1000 subjects simulated per replicate, good marginal statistics for the covariates, including 95% confidence intervals, could be obtained. Since the results were fairly invariant between the replicates (based on a small standard error of the mean), 10 replicates were judged to be sufficient to obtain precise estimates of the covariate summaries.

The Continuous and Discrete methods were applied to both real data and 27 simulated covariate data sets, as follows:

Empirical distribution of covariates (real data example)

The real data example was based on 467 subjects with seven continuous covariates (age, weight, body mass index, diastolic and systolic blood pressure, total cholesterol, fasting blood glucose) and three categorical

Table I. Parameter values for the simulated target population covariate data sets

Parameter	CONT1	CONT2
Mean (CAT = 1)	10,50,90*	90
Mean (CAT = 2)	100	100
CV (%)	30	30
Minimum	0	0
Maximum	1000	1000

Each subgroup (CAT = 1 and CAT = 2) was simulated with a separate log-normal distribution based upon the mean, coefficient of variation, and range. All parameter values were fixed except for the mean of CONT1 for subgroup CAT=1, which varied dependent upon the mode ratio, MR* (see Table II).

covariates (sex, smoking status, and diagnosis), with 2, 3, and 4 categories, respectively. Subdividing the population by unique combinations of categorical covariates created $2 \cdot 3 \cdot 4 = 24$ subgroups, with between 1 and 83 subjects per subgroup (median = 9). In this example, due to the large number of subgroups and the resulting small number of subjects in many of these subgroups, the Discrete method proved to be impractical.

Simulated distributions of covariates (simulated data example)

The simulated covariate distribution consisted of one categorical covariate with 2 levels (CAT = 1, CAT = 2), and two correlated continuous covariates (CONT1 and CONT2). Each subpopulation (CAT = 1 and CAT = 2) was simulated with a separate log-normal distribution for each continuous covariate, according to the summary statistics in Table I. Both continuous covariates had a bimodal distribution (the means of the two subgroups differed). The summary statistics for CONT2 were fixed; for CONT1, all parameters were fixed except for the mean of the CAT = 1 subgroup.

To test different simulation scenarios, three factors were varied: the percentage of patients in the CAT = 1 subgroup, the correlation between CONT1 and CONT2, and the extent of "bimodality" (overlap) of the distributions of CONT1 for the two subgroups, designated as the mode ratio, MR. The lower the value of MR, the more distinct the modes of the two subgroups were and the more the overall population distribution appeared bimodal. Therefore a low MR value (0.1) resulted in a distribution that appeared to be strongly bimodal whereas a high MR value such as 0.9 indicated a large degree of overlap between the two subgroups, leading to the appearance that the population is in fact unimodal. Table II lists the three values chosen for each factor.

Table II. Variables defining the simulation scenarios ($n = 27$) for the simulated target population covariate data sets

% (CAT = 1)	Corr	MR
10	0	0.1
25	0.45	0.5
50	0.9	0.9

The factors varied were the percentage of patients in subgroup CAT = 1, the correlation between CONT1 and CONT2, and the mode ratio, MR, which, when multiplied by 100, defined the mean of CONT1 for subgroup CAT = 1, and indicated the degree of overlap between the two subgroups. The mean of CONT1 for subgroup CAT = 2 was set to 100.

The MVND parameters for each of the possible scenarios were used to simulate 27 covariate data sets with Trial Simulator.

Categorical Covariate Coding

An analysis was completed to assess whether the order in which a categorical covariate was coded had a significant impact on the results. For the real data example, the original coding for smoking status was set at 1 for nonsmokers, 2 for former smokers, and 3 for current smokers (1/2/3). Two alternate data sets were created with the codes shuffled (2/3/1 and 3/1/2). The Continuous method was performed for the two alternate data sets and the summary statistics compared to the original data as well as to the simulated data with the initial coding.

To determine if the results of the methodology were invariant to the actual values of the codes (for example, 1/2/3 vs. 1/2/10 vs. 1/2/50), a similar analysis to the previous was performed.

RESULTS

Empirical Distribution of Covariates (Real Data Example)

Table III and Figs. 3 and 4 display the covariate summary statistics of the real data set and those generated using the Continuous Method. The mean, standard deviation, and range of the continuous covariates in the original population are maintained in the simulated population. In addition, the proportion of each value of the categorical covariates in the original population is maintained in the simulated population, showing that the mapping from continuous to discrete value calculation is appropriate.

Table III. Real Data Example: summary statistics of the categorical covariates of target population ($n = 467$) and the covariate data generated using the Discrete and Continuous methods ($n = 1000 \times 10$ replicates)

		Obs. (%)	Continuous. % (SE)	Discrete. % (SE)
Sex	1	80.3	79.9 (0.94)	80.5 (1.51)
	2	19.7	20.1 (0.94)	19.5 (1.41)
Smoking Status	1	18.6	19.0 (1.45)	18.9 (0.78)
	2	49.3	50.2 (2.13)	48.6 (1.52)
	3	32.1	30.8 (1.53)	32.5 (0.96)
Diagnosis	1	40.0	40.6 (1.43)	40.0 (0.24)
	2	38.0	37.7 (1.50)	38.1 (1.41)
	3	6.0	5.88 (0.63)	6.18 (0.45)
	4	16.0	15.8 (0.71)	15.7 (1.38)

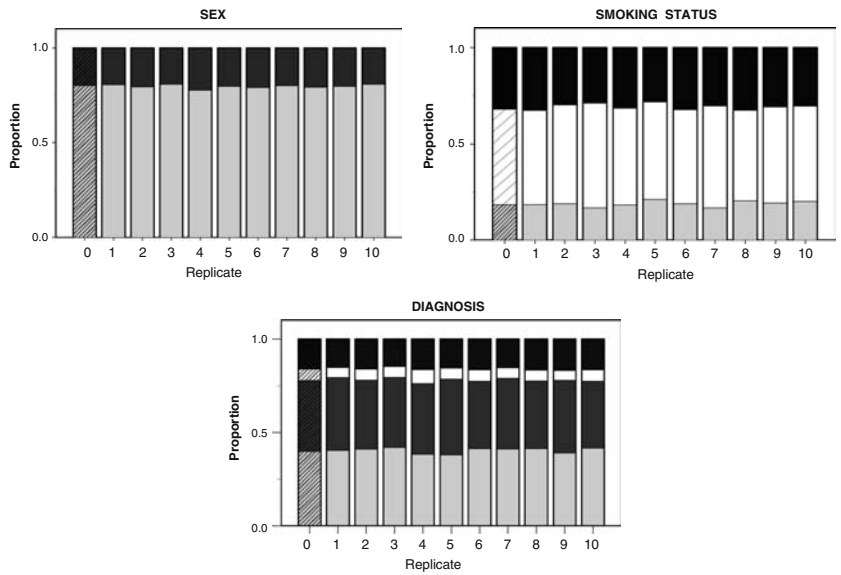

Fig. 3. Real Data Example: Proportion of patients in each level for the three categorical covariates. The leftmost bar in each plot represents the target population data ($n = 467$). Each additional bar represents one replicate of 1000 patients, generated using the Continuous method. The varying bar colors represent the proportion of subjects in each category.

The standard errors of the mean (continuous) or proportion (categorical) for each covariate (10 replicates) demonstrate high precision of the method with negligible bias.

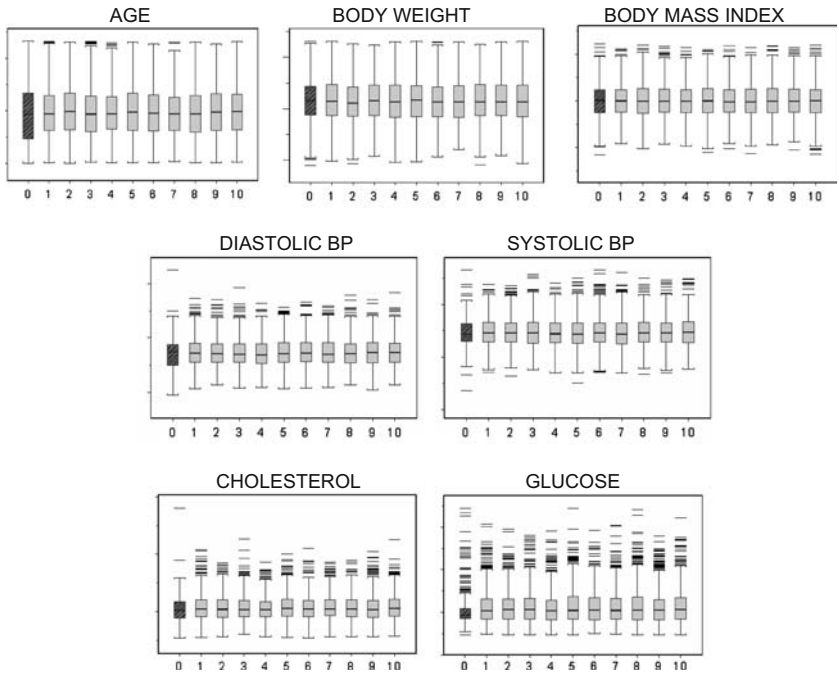

Fig. 4. Real Data Example: Box and whisker plot showing the distribution of values for the continuous covariates. The leftmost box in each plot represents the target population data ($n = 467$). Each additional box represents one replicate of 1000 patients, generated using the Continuous method. The boxes contain the 25th to 75th percentiles, the horizontal bar represents the median value, the whiskers represent the 10th to 90th percentiles, and the lines outside the whiskers represent outliers.

Due to the underlying assumptions behind the MVND, the simulated data will have a linear relationship between the covariates, which may not necessarily reflect the true shape of the relationship in the observed data. However, a test of the method's ability to preserve the correlation *coefficients* was examined by comparing the variance–covariance matrix for the observed and simulated population covariates. The maximum percent error in the covariance terms for all scenarios was 10%, indicating that the Continuous method preserved the correlation between the covariates.

Table IV displays the covariate summary statistics of the real data set and those generated using the Discrete Method. The accuracy, precision, and bias are similar to those of the Continuous method. For the real data, however, the statistics from the Discrete method using the continuous covariates are based on results from only 16 of the 24 subsets. The MVNDs for the remaining 8 subsets, containing between 1 and 7 sub-

Table IV. Real Data Example: summary statistics of the continuous covariates in the target population ($n = 467$) and the covariate data generated using the Discrete and Continuous methods ($n = 1000 \times 10$ replicates)

	Observed Mean ± SD	Continuous method Mean (SE) ± SD (SE)	Discrete method Mean (SE) ± SD (SE)
Age	68.7 ± 8.2	69.7 (0.23) ± 7.20 (0.15)	69.9 (0.21) ± 4.98 (0.14)
Weight	72.6 ± 12.8	74.0 (0.27) ± 10.8 (0.19)	70.5 (0.44) ± 7.19 (0.34)
BMI	25.8 ± 3.5	26.1 (0.04) ± 2.92 (0.07)	25.8 (0.08) ± 2.41 (0.09)
Cholesterol	205.5 ± 44.0	214.0 (1.50) ± 43.7 (1.32)	212.2 (2.29) ± 30.95 (1.19)
Diastolic BP*	77.9 ± 11.0	79.4 (0.36) ± 10.2 (0.16)	79.4 (0.36) ± 10.2 (0.16)
Systolic BP*	146.1 ± 18.4	148.1 (0.36) ± 17.1 (0.42)	147.0 (1.06) ± 12.5 (0.45)
Glucose	5.97 ± 1.91	6.70 (0.07) ± 2.36 (0.09)	6.48 (0.11) ± 1.39 (0.07)

* BP: Blood Pressure

jects, failed to simulate reasonable values for the continuous covariates. Additional analyses were performed to determine the causality of this failure; it was found that if there are data from less than $N + 1$ subjects in a subgroup (where N is the number of covariates in the MVND), the variance–covariance matrix that is generated will be singular. This is clearly a potential disadvantage of the Discrete method.

Simulated Distributions of Covariates (Simulated Data Example)

Figure 5 displays the proportion of patients in the (CAT = 1) subgroup, and shows that both the Continuous and Discrete methods generate the expected proportion of patients in each subgroup.

As shown in Fig. 6a, for the situations in which MR equals 0.1, indicating little overlap between the subgroups, the Continuous method fails to describe the distribution of CONT1 in the two subgroups. However, as MR increases, the subgroup distributions begin to overlap, masking the bimodal characteristics of the distribution, and the Continuous method begins to describe the original distributions more accurately. When the overlap is substantial (MR = 0.9, Fig. 6c), the distribution of CONT1 generated by the Continuous method matches the simulated covariate data distribution almost exactly. For all values of MR, the Discrete method adequately distinguishes the distributions of the individual subgroups for CONT1; this is expected since each subgroup is simulated separately.

The percent prediction errors (%PE) in the summary statistics of CONT1 for the Continuous method are shown in Fig. 7. %PE is calculated from 100· (predicted-true)/true, where true is the true mean of CONT1, and "predicted" is the mean of CONT1 in the covariate data

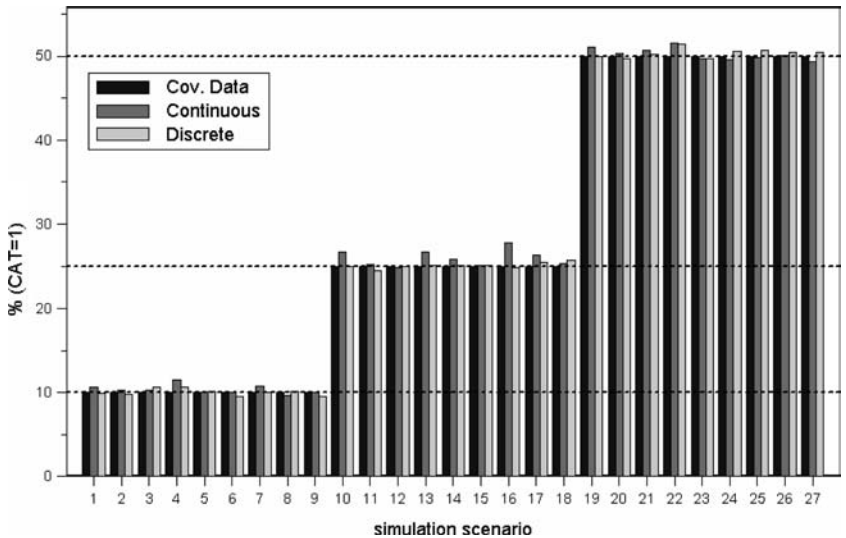

Fig. 5. Simulated Data: Bar chart showing the percentage of patients in the subgroup (CAT = 1) for the target population covariate data and the covariate data generated by the Continuous and Discrete methods. There should be 10, 25, and 50% in the (CAT = 1) subgroup, respectively, for each set of 9 scenarios.

simulated by the Continuous method. For the whole population, the Continuous method reliably simulates covariates with mean and coefficient of variation close to the true values. The %PE is negligible, and is relatively independent of MR and number of patients in the subgroups, with only a slight negative %PE for both the mean and CV at a MR value of 0.1. However, for the individual subgroup summary statistics, the %PE is highly dependent upon both MR and the percentage of patients in that subgroup. For MR = 0.1, the Continuous method results overestimate the mean and CV of CAT = 1 (as shown by a large positive %PE); as the percentage of patients in this subgroup increases, however, the error decreases. As MR increases, the errors approach zero for both mean and CV in the subpopulations. For the Discrete method, there are negligible errors in the mean and SD for the subgroups and for the whole population, which are independent of the values of MR or the percentage of patients in each subgroup (results not shown).

Figure 8 shows the correlation between CONT1 and CONT2 for the true simulated covariate data, and for the Continuous and Discrete method results. Comparing the plots of the simulated covariate data and the Continuous method results indicates that for a MR value of 0.1, the continuous method fails to capture the relationship between CONT1 and CONT2, but adequately captures the correlation for larger ratios. For all

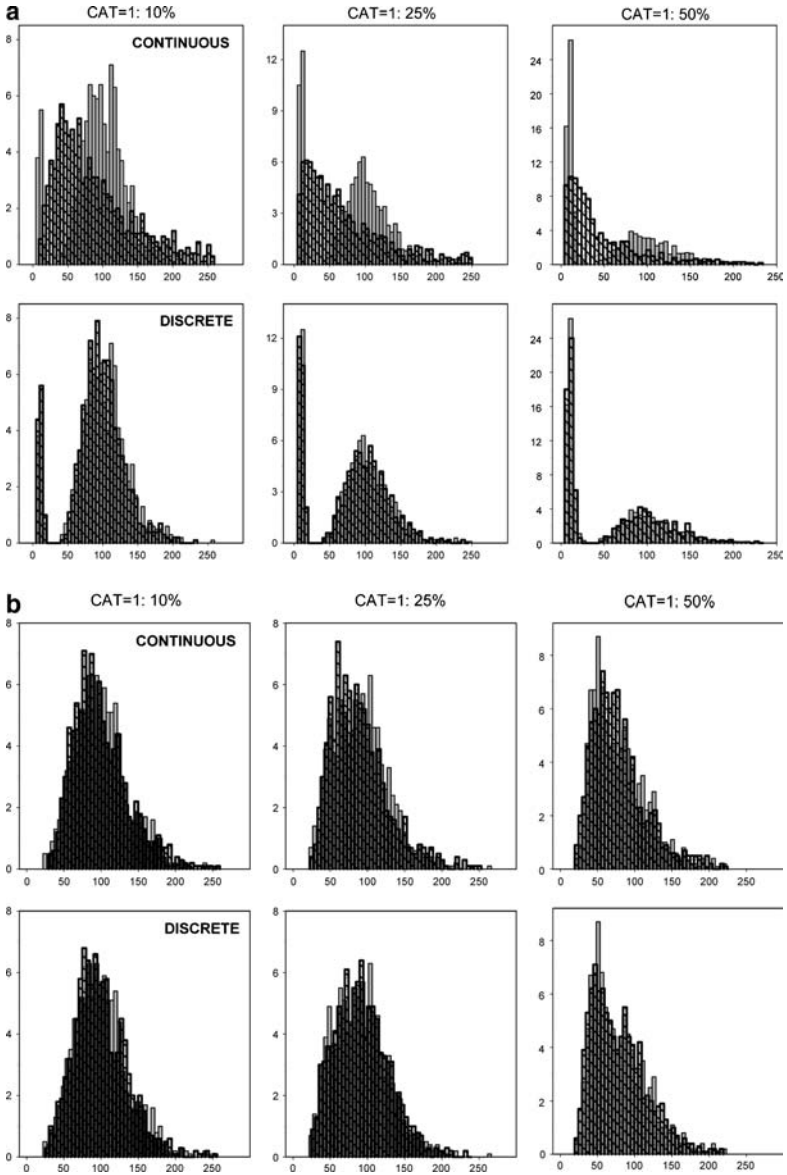

Fig. 6. (a) Simulated Data: Population distribution of CONT1 for MR = 0.1. The target population covariate data (gray bars) is overlaid with the Continuous method results (top) and Discrete method results (bottom), shown as transparent bars. Only the scenarios for correlation = 0 between CONT1 and CONT2 are shown, but the plots for correlations of 0.45 and 0.9 look similar. (b) Simulated Data: Population distribution of CONT1 for MR = 0.5. (c) Simulated Data: Population distribution of CONT1 for MR = 0.9.

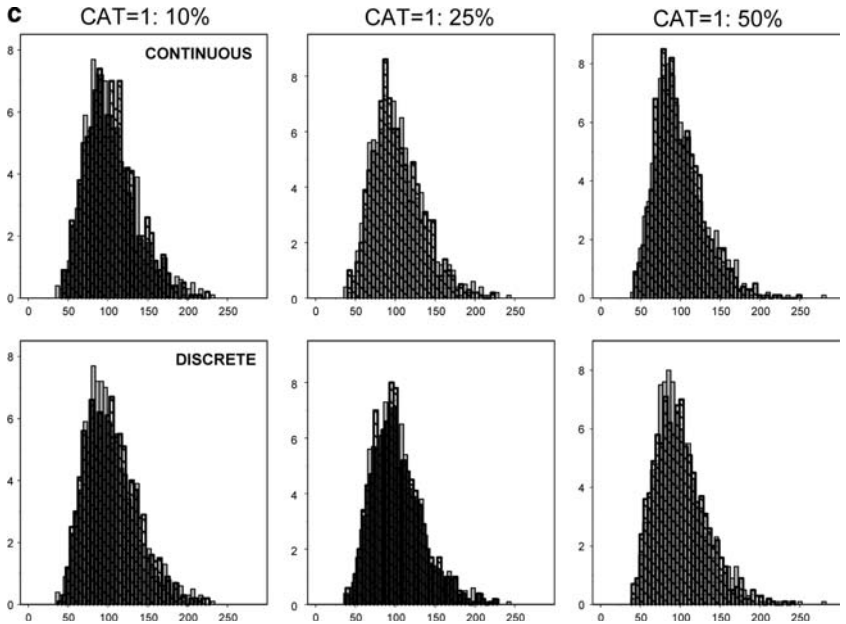

Fig. 6. Continued.

ratios, the Discrete method accurately captures the correlation between CONT1 and CONT2.

Figure 9 shows the %PE for the correlation between CONT1 and CONT2. For MR = 0.1, the Continuous method underestimates the correlation between CONT1 and CONT2 (large negative %PE) for both the subpopulations and the whole population. As MR increases, the absolute errors decrease in the subpopulations, and are negligible in the whole population. There does not appear to be a relationship between percent of patients in each subgroup and the correlation. For the Discrete method, there are negligible errors in correlation for the subgroups or for the whole population for all ratios and proportion of patients in each subgroup (results not shown).

Categorical Covariate Coding

Shuffling the order of the coding by definition altered P_i (see Eq. 1), but also transformed the mean and standard deviation for smoking status, and thus the CrVs. Paired with the new variance–covariance matrix (the row and column for smoking status was altered), the new parameters for the MVND compensated for the shuffled coding. Therefore, the order of categorical

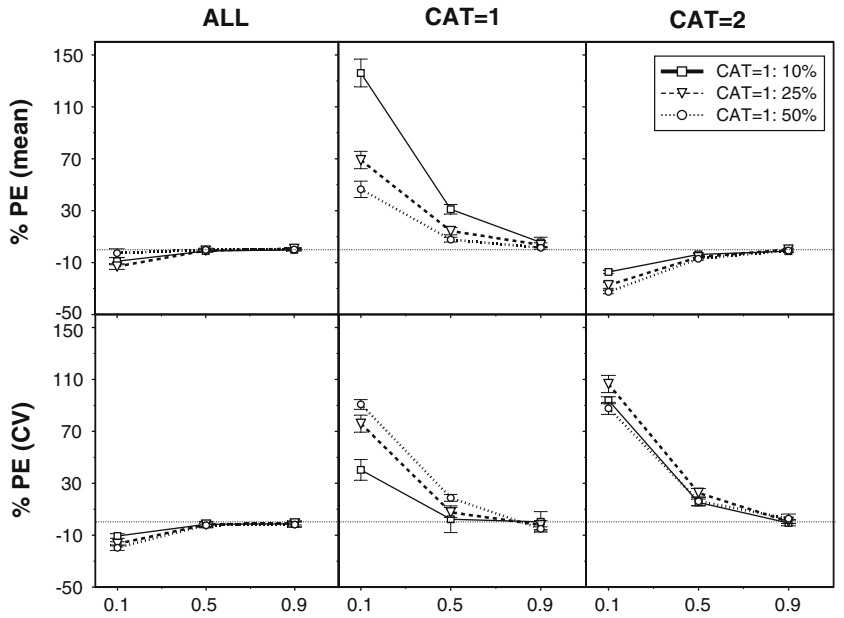

Fig. 7. Simulated Data: Percent error in mean (top row) and CV (bottom row) of CONT1 for the values generated by the Continuous method, as a function of mode ratio and % (CAT = 1).

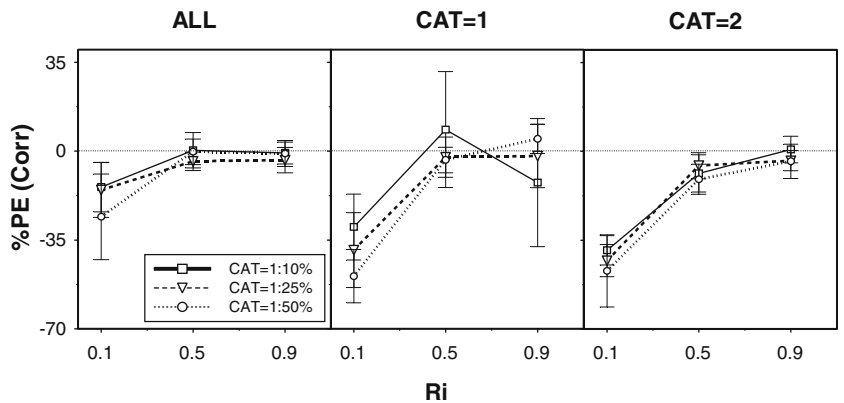

Fig. 8. Simulated Data: Correlation between CONT1 and CONT2 for CAT = 1 (light gray) and CAT = 2 (dark gray). Only the scenarios with 50% of patients in each subgroup are shown, but the plots for 10% and 25% in the CAT = 1 subgroup look similar.

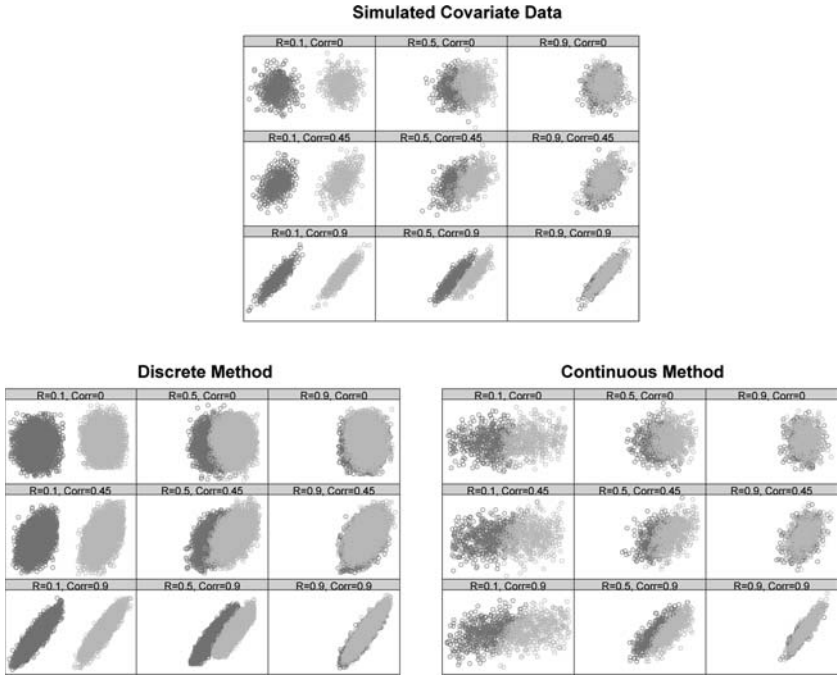

Fig. 9. Simulated Data: Percent error in correlation between CONT1 and CONT2 for the values generated by the Continuous method, as a function of mode ratio and % (CAT = 1).

covariate coding had no effect upon the simulated results. The summary statistics of the continuous covariates, and the proportion of each value of the categorical covariates in the original population, including smoking status, were nearly identical to that of the original observed population.

While the value of the codes (1/2/3 vs. 1/2/50) did impact the simulated results when the third value became large, this analysis was really not necessary; should codes such as the latter appear in a data set, the third value could be simply recoded (i.e., all values of 50 changed to 3) for the creation of the MVND, and then transformed back (to 50) in the simulated data set.

DISCUSSION

As demonstrated by the real data example, both the Continuous method and Discrete methods generate accurate summary statistics for the covariates of the target population. The mean, standard deviation, and range of the continuous covariates in the target population, and the

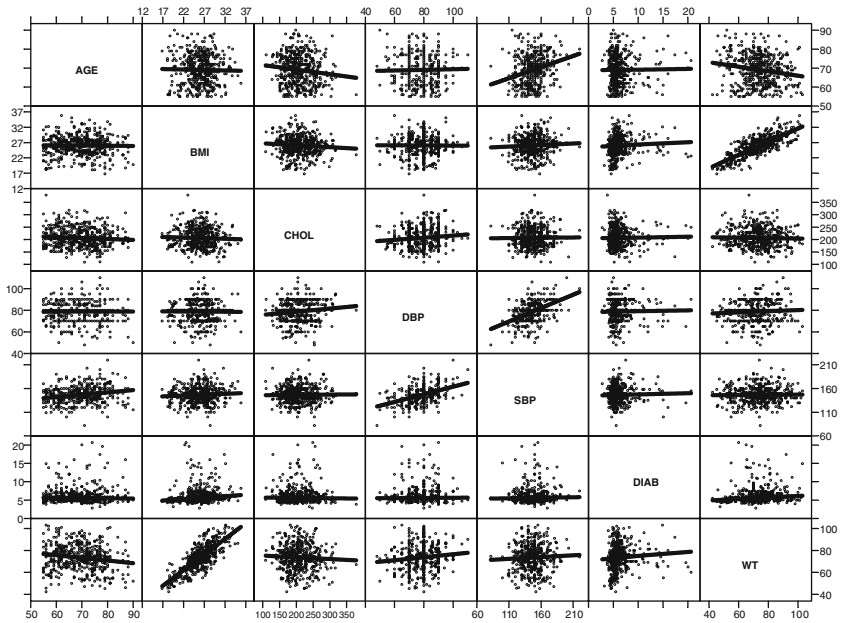

Fig. 10. Scatterplot matrix for observed continuous covariates in real data set. Lines represent Loess smooths of the individual plots, and indicate that the relationships between the covariates are relatively linear; thus, it is appropriate to enter all covariates into the MVND.

proportion of patients with each of the categorical covariate values, are maintained in the population simulated using either method. In addition, the standard error of the 10 replicates shows high precision of both methods, while the prediction error indicates negligible bias.

However, despite the fact that the Discrete method appears to generate the proper values for the target population summary statistics, these results are misleading because these statistics were calculated from only 2/3 (16 of 24) of the unique combinations of categorical covariate values. The remaining subsets contained fewer than 8 subjects (one plus the number of continuous covariates); the amount of data in each subset was thus inadequate to obtain a nonsingular variance–covariance matrix between the covariates. As the number of patients in each subgroup increases, the covariate variance–covariance matrix derived from these data becomes more precisely estimated. Consequently the summary statistics of the simulated data more closely match the original statistics, indicating that the model is representative of these patient subsets in the clinical trial. By not dividing the patients into subsets, the Continuous method ensures that there will be sufficient data to generate the MVND; in addition, since each

subset requires calculation of summary statistics and creation of a separate MVND, the number of analyses that must be performed is reduced.

In the simulated covariate data case, both the Continuous and Discrete method results match the summary statistics of the whole population, with low error in the means and CVs, as well as the expected proportion of patients in each subgroup. However, whereas the Discrete method adequately recreates the distribution of the CONT1 values for all values of MR, the Continuous method does not. Because the Continuous method assumes a unimodal distribution for the covariate in the whole population but in fact a bimodal distribution for a covariate exists, the Continuous method is not able to recreate this distribution when MR is low. However, as the subgroups overlap (i.e., the value of MR increases), the bimodal characteristics of the distribution of CONT1 become obscured, and the overall population distribution appears unimodal. Our example suggests that when MR is greater than or equal to 0.5, the performance of the Continuous method is adequate.

The Continuous method has limited utility for populations when the subgroup means are very different; however, to our knowledge, there are few clinically relevant examples in which such a low value of MR might be seen. Two possible examples are testosterone/estrogen concentration in males versus females (17), or carboxyhemoglobin concentration in smokers versus nonsmokers (18). For most commonly used covariates, such as weight and age, the means of these covariates are usually quite similar between subjects with different sexes, races, or smoking status, with an MR value of perhaps 0.8 or higher. However, even when MR is low between subgroups, the Continuous method may still be utilized when the Discrete method is challenged by inadequate numbers of subjects in subgroups in empirically derived datasets. Rather than completely subdividing the population into each unique combination of categorical covariate value, the subgroups with a low value of MR may be separated out and the Continuous method could then be applied to describe the remaining covariates.

Because a MVND simulates covariates whose pairs are linearly related, the results may not represent the true shape of the covariate relationships in the underlying empirical distribution. It is therefore recommended that a scatterplot matrix be created for the covariates, as shown in Fig. 10 for the real data example. This will allow a visual examination of the relationship between the covariates, and the opportunity to identify any covariate that shows a significantly nonlinear association with another. For example, if covA and covB are related nonlinearly, there are methods that can be carried out prior to creating the MVND to correct this. First, one might attempt to transform the covariates in some way to linearize their relationship (e.g., for a relationship represented by a power

model, the log transformed covariates are linearly correlated). Alternately, the relationship between covA and covB may be modeled with both fixed and random effects to obtain a mathematical expression; covB could then be excluded when estimating the parameters for the MVND. After sampling values for the covariates from the MVND, for each value of covA an estimate of covB can be obtained from the quantitative relationship between the two covariates. Similarly, this method could be used to reconcile two covariates that are linearly dependent, which may cause the variance–covariance matrix to be singular if both are left in the analysis.

The Continuous method can be valuable when the original covariate data from which a MVND is derived does not exactly match the desired target patient population for the clinical trial simulation. Although the individual covariates themselves may be the same, it may be desirable to simulate virtual patients in a different age group, or perhaps to explore the outcome of the model given a larger percentage of smokers than in the original data set. To do this, one simply has to adjust the inclusion–exclusion criteria for the simulation study without changing the MVND. For example, if it was desired to have 50% males and 50% females in a 100 subject study, the criteria would be set such that once 50 males were selected, all subsequent covariate vectors from male subjects would be excluded. Similarly, if older patients were desired, covariate vectors with patients below a certain age would be excluded. Once all subjects are selected, the demographics of this final population should still match that of the original patient population, due to the fact that the subjects were selected from the MVND created from these data. It is assumed that the MVND created from the original population represents the inherent interrelationships between the selected covariates for a typical subject population. These relationships may be borrowed from other similar subject populations, even if the overall demographics (mean age, percentage of smokers, etc.) were different.

We hope in the future to investigate this method for more elaborate covariate distribution models (for example, time-varying covariates).

Compared to the Discrete method, the Continuous method has a number of benefits that result from analyzing the whole population instead of small subsets. Including a large amount of data in the creation of the variance–covariance matrix enhances its stability and, as a consequence, the reliability of the generated covariate combinations. In addition, by allowing all covariates to be described by a single MVND (rather than one for each unique combination of categorical covariates), the number of analyses that must be performed is reduced, increasing efficiency. With the exception of the rare instance of a low mode ratio between continuous distributions, the Continuous method appears to efficiently generate

unbiased, precise covariates for the purposes of simulating virtual patient covariate vectors in a clinical trial simulation.

REFERENCES

1. N. H. G. Holford, M. Hale, H. C. Ko, J.-L. Steimer, and C. C. Peck (eds.). P. Bonate, W. R. Gillespie, T. Ludden, D. B. Rubin, L. B. Sheiner, and D. Stanski (contributors). *Simulation in Drug Development: Good Practices*. http://cdds.georgetown.edu/research/sddgp723.html
2. P. L. Bonate. Clinical trial simulation in drug development. *Pharm. Res.* **17**:252–256 (2000).
3. N. H. G. Holford, J. Monteleone, H. Kimko, and C. Peck. Simulation of clinical trials. *Annu. Rev. Pharmacol. Toxicol.* **40**:209–234 (2000).
4. S. Chabaud, P. Girard, P. Nony, and J. P. Boissel. Clinical trial simulation using therapeutic effect modeling: Application to ivabradine efficacy in patients with angina pectoris. *J. Pharmacokinet. Pharmacodyn.* **29(4)**:339–363 (2002).
5. H. J. M. Lemmens, D. R. Wada, C. Munera, A. Eltahtawy, and D. R. Stanski. Enriched analgesic efficacy studies: An assessment by clinical trial simulation. *Contemp. Clin. Trials* **27(2)**:165–173 (2006).
6. C. Veyrat-Follet, R. Bruno, and R. Olivares. Clinical trial simulation of docetaxel in patients with cancer as a tool for dosage optimization. *Clin. Pharmacol. Ther.* **68**:677–678 (2000).
7. H. Kastrissios, S. Rohatagi, J. Moberly, K. Truitt, Y. Gao, D. R. Wada, M. Takahashi, K. Kawabata, and D. Salazar. Development of a predictive pharmacokinetic model for a novel cyclooxygenase-2 inhibitor. *J. Clin. Pharmacol.* **46(5)**:537–548 (2006).
8. K. G. Kowalski and M. M. Hutmacher. Design evaluation for a population pharmacokinetic study using clinical trial simulation: A case study. *Stat. Med.* **20**:75–91 (2001).
9. D. R. Mould. Defining covariate distribution models for clinical trial simulation. In H. C. Kimko and S. B. Duffull (eds.), *Simulation for Designing Clinical Trials: A Pharmacokinetic– Pharmacodynamic Modeling Perspective*, Marcel Dekker, New York, 2003, pp. 31–53.
10. M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*, John Wiley & Sons, Inc., New York, 1993, pp. 102–105.
11. S. L. Beal and L. B. Sheiner (eds.). NONMEM Users Guides, Icon Development Solutions, Ellicott City, MD (1989–98).
12. B. Schmeiser. Advanced input modeling for simulation experimentation. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans (eds.), *Proceedings of the 1999 Winter Simulation Conference*, 1999, pp. 110–115.
13. M. Kaut and S. W. Wallace. Evaluation of scenario-generation methods for stochastic programming (2003). http://citeseer.ist.psu.edu/cache/papers/cs/29430/http:zSzzSzwww.iot.ntnu.nozSz mkautzSzCV_and_studyzSzSG_evaluation.pdf/kaut03evaluation.pdf
14. S. Ghosh and S. G. Henderson. Chessboard distributions and random vectors with specified marginals and covariance matrix. Working paper, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor (2000).
15. M. C. Cario and B. L. Nelson. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical Report, Department of Industrial Engineering and 35 Management Sciences, Northwestern University, Evanston, Illinois (1997).
16. L. Lapin. *Probability and Statistics for Modern Engineering* (2nd edn), PWS-KENT Publishing Company, Boston, 1983, pp. 215–217.
17. R. H. Williams and P. R. Larson (eds.), *Williams Textbook of Endocrinology* (10th edn), W.B. Saunders Co., Illinois, 2002, pp. 622, 726.
18. P. McDonough and R. Moffatt. Smoking-induced elevations in blood carboxyhaemoglobin levels. Effect on maximal oxygen uptake. *Sports Med.* **27**:275–283 (1999).