A Novel Method for Simulation of Correlated Continuous and Categorical Variables Using A Single Multivariate Distribution

S. Tannenbaum (1,2), Nick Holford (2,4), H. Lee (2,3), C. Peck (2), D. Mould (5)

(1) Novartis Pharmaceuticals Corp., East Hanover, NJ, USA, (2) Center for Drug Development Science, Washington, DC, USA, (3) University of Pittsburgh, Pittsburgh, PA, USA, (4) University of Auckland, Auckland, NZ, (5) Projections Research, Inc., Phoenixville, PA, USA

Introduction

Objective

To test a novel method, which treats categorical covariates as continuous, for generating virtual patients for clinical trial simulation. This method will be compared to the standard method for generating covariate vectors, which involves generating distributions of continuous covariates for each unique combination of categorical covariates.

Covariate Distribution Model

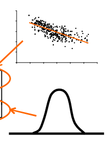
Generates covariate vectors representing each virtual patient in a clinical trial simulation

- Covariate combinations must be realistic and reasonable
- Covariate values are often constrained to pre-specified target population demographics

Sampling from a Multivariate Normal Distribution (MVND)

- Complete patient covariate vectors are sampled from a multivariate probability density function
- The simulation platform creates this function given:
  - o Central tendency (mean) of each covariate
  - o Covariate variance-covariance matrix (VCVM)
    - – The diagonal elements of the VCVM, shown below, are the variance values for each individual covariate
    - – The off-diagonal elements are the covariance values indicating the relationship between each pair of covariates

$$\begin{array}{cc} & X \quad\quad Y \\ \begin{array}{c} X \\ Y \end{array} & \begin{bmatrix} \mathrm{var}(X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(X,Y) & \mathrm{var}(Y) \end{bmatrix} \end{array}$$



- – VCVM, and mean, low, and high values for each covariate are entered into the simulation platform for generation of virtual patients
- Benefit: unique, reasonable, and realistic patient covariate vectors will be created
- **Limitation: requires all covariates to be continuous and have the same distribution**

How can one incorporate categorical covariates into a covariate distribution model using a multivariate normal distribution?

Standard method: Discrete Method

For each unique combination of categorical covariates, sample continuous covariates from separate MVNDs

- Stratification of patients into subgroups leads to reduced numbers of patients available in each category for evaluation (i.e., insufficient data to build a representative model)
- Can be cumbersome to implement when there are many categories

Novel method: Continuous Method

Sample from a single MVND created by treating **all** covariates as continuous

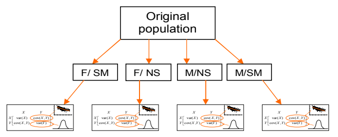
- Not necessary to stratify patients into subgroups
  - o Analyzing a whole population instead of small subsets increases the stability of the joint function and the reliability of the generated covariate combinations
  - o The number of analyses that must be performed is reduced

Methods

Discrete Method (DM)

- Categorical covariates are sampled from their individual distributions
- Continuous covariates are then generated from the proper subgroup's MVND

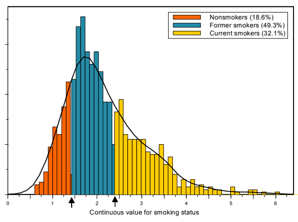
Ex: 2 categorical covariates: sex (M/F) and smoking (NS/SM)



Continuous Method (CM)

- All covariates are treated as continuous and log-normally distributed
  - o the resultant single MVND is used to generate complete patient covariate vectors.
- Categorical covariates (e.g., X) will have continuous values
  - o cutoff values to assign the categorical levels are defined as the inverse of the lognormal cumulative distribution of X: mean(lnX), sd(lnX), and cumulative probability P ($X \le X_i$)

Example: cutoff values for smoking status (3 categories)
- Solid line represents continuous probability distribution curves
- Histogram represents empirical distribution
- Arrow on X-axis represent cutoff values



Method qualification

The CM and DM were applied to real and simulated data sets to compare their abilities to generate matching virtual patient distributions.

Empirical Distribution of Covariates (Real Data Example)

- n=467
- 7 continuous covariates
  - o age, weight, body mass index, diastolic and systolic blood pressure, total cholesterol, fasting blood glucose
- 3 categorical covariates
  - o sex (2 categories), smoking status (3 categories), diagnosis (4 categories)

Simulated Distributions of Covariates (Simulated Data Example)

- One categorical covariate with 2 levels (CAT=1, CAT=2)
- Two continuous covariates (CONT1 and CONT2)
  - o Each subpopulation (CAT=1 and CAT=2) simulated with a separate log-normal distribution for each continuous covariate

- Fixed parameters

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

$$\text{* Mode Ratio} = \frac{\text{mean of CONT1(CAT = 1)}}{\text{mean of CONT1(CAT = 2)}}$$

- Simulation Scenarios (n=27)

% (CAT=1) *	Corr**	Mode Ratio***
10	0	0.1
25	0.45	0.5
50	0.9	0.9

* Percentage of patients in subgroup CAT=1
** Correlation between CONT1 and CONT2
*** Low ratio indicates completely separate subgroups, high ratio indicates overlapping subgroups

- 10 replicates of 1000 virtual patients were simulated for each scenario

Qualification Steps

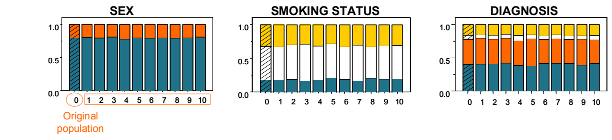
- 1000 subjects were simulated using both the DM and CM
  - o simulation was replicated 10 times
- Metrics compared to "observed" (real or simulated) data
  - o population summary statistics
  - o distributions of continuous covariates
  - o proportions of categorical covariate values
  - o correlation between CONT1 and CONT2 (simulated data set only)

Results

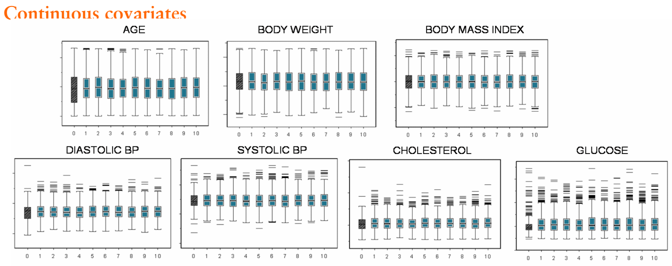
Empirical Distribution of Covariates

Figures show results for CM only- DM results are approximately the same

Categorical covariates



Continuous covariates



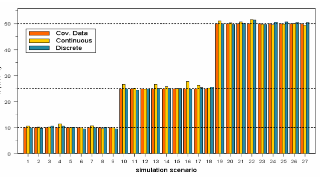
- For both the CM and DM, compared to "observed" data:
  - o mean, standard deviation, and range of the continuous covariates, and proportion of each value of the categorical covariates is maintained
    - – Positive results for CM demonstrate that the mapping from discrete to continuous then back to discrete is appropriate
  - o The standard errors of the mean (continuous) or proportion (categorical) for each covariate (10 replicates) demonstrate high precision of the method with negligible bias

- The outcome from the DM is based on results from only **16 of the 24 subsets**.
  - o The remaining 8 subsets contained between 1 and 7 subjects
  - o **If there are data from less than N+1 subjects in a subgroup (where N is the number of covariates in the MVND), the VCVM will be singular**

Simulated Distribution of Covariates

Categorical covariates

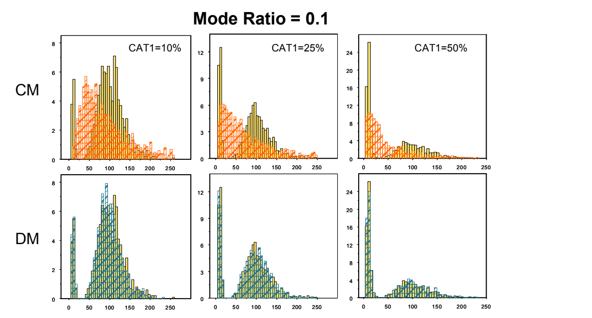
Percentage of patients in the subgroup (CAT=1) for the observed data, CM, and DM. There should be 10%, 25%, and 50% in the (CAT=1) subgroup, respectively, for each set of 9 scenarios

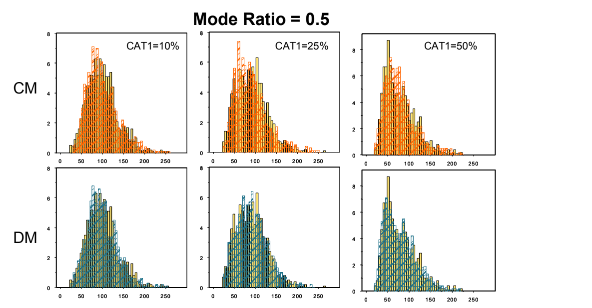


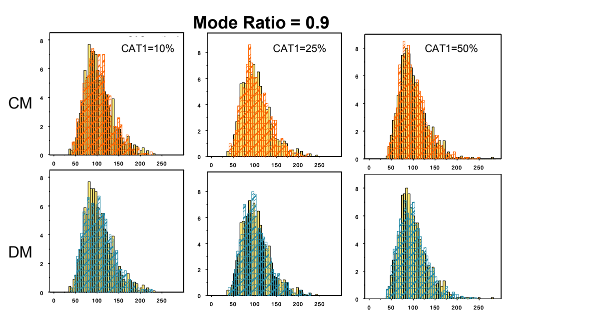
Categorical covariates

Only scenarios for correlation = 0 between CONT1 and CONT2 are shown (plots for correlations of 0.45 and 0.9 look similar)

- *"observed" covariate data (yellow bars), CM (orange), DM (blue)*







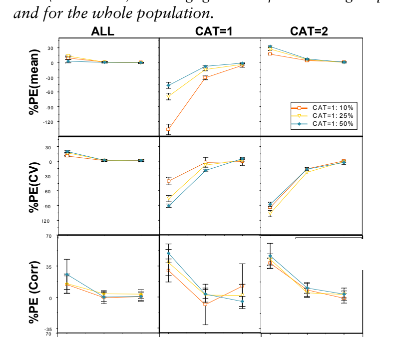
Simulated Distribution of Covariates

Correlations

Percent prediction errors (%PE) in the summary statistics of CONT1 (shown for CM)

%PE = 100·(predicted-true)/true

DM (not shown) had negligible PE for the subgroups and for the whole population.



Full population: CM reliably simulates covariates with mean and coefficient of variation close to the true values.

Individual subgroup summary statistics: %PE is dependent upon both MR and the percentage of patients in that subgroup

- as the percentage of patients in CAT=1 increases, the error decreases.
- as MR increases, the errors approach zero for both mean and CV in the subpopulations.

Negligible errors for DM, independent of MR or the percentage of patients in each subgroup

Discussion

CM and DM generate accurate summary statistics for the covariates of the target population for both real and simulated data

DM results are misleading for real data

- Appears to generate the proper values for the target population summary
- the amount of data in 8/24 subsets was inadequate to obtain a non-singular VCVM

DM adequately recreates the shape of the bimodal distribution for CONT1 for all values of MR

CM assumes a unimodal distribution for the covariates in the whole population

- – As MR increases (subgroups overlap), the bimodal characteristics become obscured
- – CM is successful when overall population distribution appears unimodal
  - few clinically relevant examples in which a very low value of MR might be seen

Hybrid CM/DM may be utilized when there are inadequate numbers of subjects in subgroups

- – Rather than completely subdividing the population, the subgroups with a low value of MR may be separated out
- – CM could then be applied to describe the remaining covariates

CM can simulate novel patient populations for clinical trial simulation

- – Adjust the inclusion-exclusion criteria for the simulation study without changing the MVND
- – Assumption: the MVND from the original population represents the inherent interrelationships between the covariates, even if the overall demographics (mean age, percentage of smokers, etc.) were different

Conclusion

CM has a number of benefits that result from analyzing the whole population instead of small subsets

- – Large amount of data in the creation of the VCVM enhances its stability and, as a consequence, the reliability of the generated covariate combinations.
- – 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 MR, the CM appears to efficiently generate unbiased, precise covariates for the purposes of simulating virtual patient covariate vectors in a clinical trial simulation.

PAGE (Population Approach Group Europe)
Pamplona, June 2005

NOVARTIS