Conditional distribution modeling for covariates simulation using classification and regression trees methods

Giovanni Smania[1], E. Niclas Jonsson[1]
[1]Pharmetheus, Uppsala, Sweden

Pharmetheus

Background

- To obtain realistic scenarios in clinical trial simulation, simulated subjects must be representative of the target population.
- Common ways of generating virtual subjects are based upon bootstrap procedures or multivariate normal distributions (MVND) [1]. We recently investigated an alternative method based on conditional distributions (CD), which used predictive mean matching (PMM) as underlying prediction model [2].
- Previous studies have shown that, in the context of missing data imputation, CD with classification and regression trees (CD-CART) outperformed CD with PMM (CD-PMM) when strong interactions/nonlinear effects among the variables exist [3].

Objectives

- To investigate the operating characteristics of CD when used to simulate covariates distributions based on CART methods
- To compare the performance of CD-CART vs. CD-PMM.

Methods

- 233 healthy subjects and 706 patients from a real drug development dataset were used (Table 1 and 2).

Table 1. *Summary statistics of continuous covariates.*

Covariate	Age	WT	SCR	CRCL
Mean (SD)	46.4 (12.4)	88.5 (20.1)	78.5 (16.2)	124 (34.3)
Median (range)	47.0 (18.0-77.0)	86.6 (46.3-172)	78.7 (41.5-133)	120 (47.0-282)

Table 2. *Summary statistics of categorical covariates.*

Sex	Race
Male n=534 (56.9%), Female n=405 (43.1%)	White n= 737 (78.5%), Black n=179 (19.1%) Asian n=19 (n=2%), American Indian n=2 (0.2%), Other n=2 (0.2%)

- CD-PMM and CD-CART were implemented using the R package mice [4] (Figure 4). N=30 datasets were simulated. The methods were evaluated based on the observed dataset (internal evaluation) as well as on their ability to predict an older population (extrapolation).

Results

Internal evaluation

- CD-CART had generally a lower absolute bias and RMSE for the mean, median, SD and range of continuous covariates (Figure 1A) and the proportion of categorical covariates (Figure 1B).
- Absolute bias in the variance-covariance matrix was comparable between the two methods (Figure 1C).
- CD-CART allowed to considerably increase the precision of the correlation structure (Figure 1C, RMSE gains up to 11%), particularly in case of highly non-linearly related covariates (Figure 2).

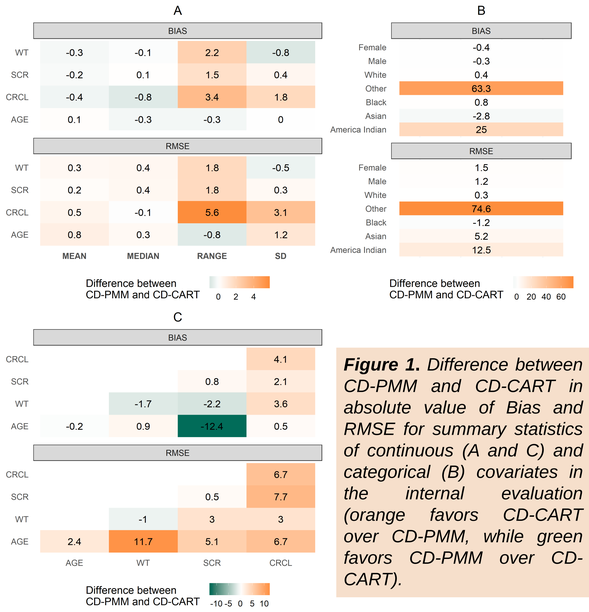
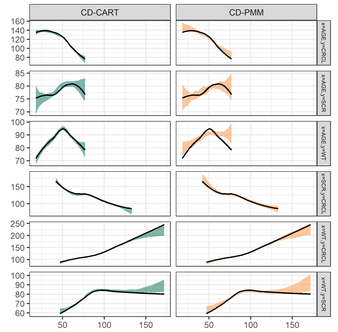


Figure 1. *Difference between CD-PMM and CD-CART in absolute value of Bias and RMSE for summary statistics of continuous (A and C) and categorical (B) covariates in the internal evaluation (orange favors CD-CART over CD-PMM, while green favors CD-PMM over CD-CART).*

Extrapolation

- CD-PMM had a slightly lower accuracy and precision in means and medians, but performed remarkably better for SD (Figure 3A).
- Absolute bias and RMSE in the proportion of males/females were generally higher for CD-PMM vs. CD-CART (Figure 3B).
- CD-CART provided better estimates of the off-diagonal terms of the variance-covariance matrix except for the WT~CRCL and AGE~CRCL relationships (Figure 3C).

Figure 2. *Visual predictive check of the relationship between pairs of continuous covariates: the black line represents a loess regressor through the original data set, while the shaded area depicts the 80% confidence interval of the loess regressor fitted to each of the 30 replicates.*



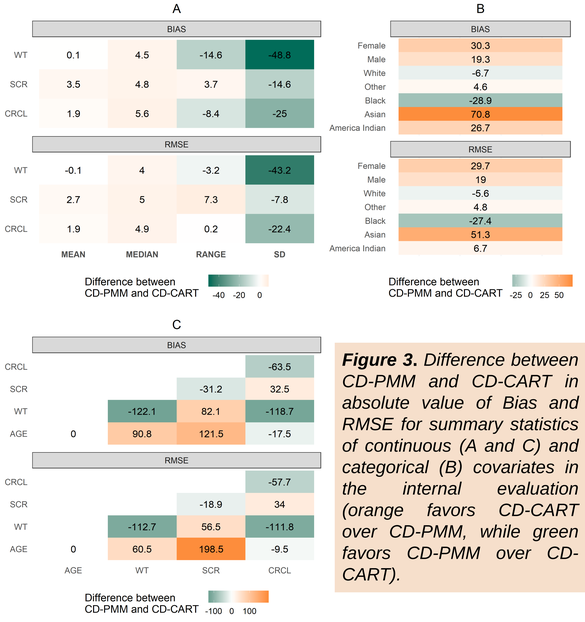


Figure 3. *Difference between CD-PMM and CD-CART in absolute value of Bias and RMSE for summary statistics of continuous (A and C) and categorical (B) covariates in the internal evaluation (orange favors CD-CART over CD-PMM, while green favors CD-PMM over CD-CART).*

References

1. Teutonico D et al. Pharm Res. 32(10), 3228-3237 (2015).
2. Smania G, Jonsson EN. CPT Pharmacometrics Syst. Pharmacol., 10: 330-339 (2021).
3. Doove LL et al. Comput Stat Data Anal. 72, 92–104 (2014)
4. van Buuren S, Groothuis-Oudshoorn K. J Stat Soft, 45(3), 1-67 (2011).

Conclusions

- CD-CART has improved operating characteristics compared to CD-PMM in the internal evaluation.
- In the extrapolation setting, CD-PMM outperformed CD-CART in terms of summary statistics of continuous covariates, while categorical covariates were better predicted by CD-CART.
- CD-CART appears to be a promising alternative to CD-PMM when dealing with covariate distributions characterized by strong non-linearities and/or interactions effects across covariates.

Using *mice* for covariates simulation

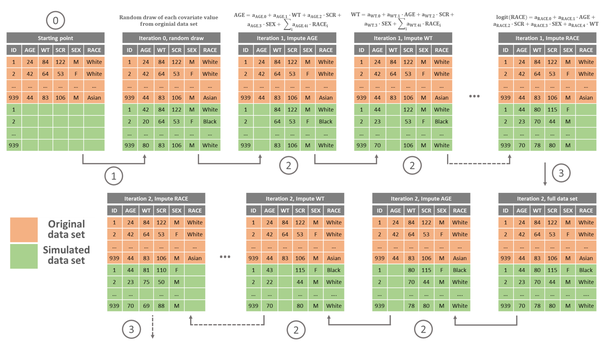


Figure 4. *Illustrative example depicting how the mice package is tweaked to simulate covariate distributions, using a standard linear regression model. The starting point is represented by a dummy data set with the same rows as the original data set and all covariates assumed to be missing (green table), which is attached to the original data set identified by the orange color (0). At iteration 0 all the missing data are first imputed by drawing a random sample from the original data set on a covariate-by-covariate basis (1). During iteration 1, missing covariates are sequentially imputed: the covariate under question is modeled conditionally on all the others and the model prediction is used as imputation (2). When all the covariates have been imputed with the respective imputation model, a new full data set is obtained (3), which is then used as starting point for iteration 2. This iterative process goes on until convergence is reached.*

Contact

giovanni.smania@pharmetheus.com
www.Pharmetheus.com