Conditional distribution modeling as an alternative method for covariates simulation: comparison with joint multivariate normal and bootstrap techniques

Giovanni Smania[1]; E. Niclas Jonsson[1]

[1] Pharmetheus AB, Uppsala, Sweden.

Corresponding author: Giovanni Smania, giovanni.smania@pharmetheus.com

Uppsala Science Park

Dag Hammarskjölds väg 36b

75237 Uppsala, Sweden

Tel: +46(0)18 513 328

Key words: *Covariate distribution modeling, multivariate distribution, conditional distribution, predictive mean matching, clinical trial simulation*

CONFLICT OF INTEREST

The authors declare no conflicts of interested.

FUNDING

This work was supported by Pharmetheus AB.

Abstract

Clinical trial simulation (CTS) is a valuable tool in drug development. To obtain realistic scenarios, the subjects included in the CTS must be representative of the target population. Common ways of generating virtual subjects are based upon bootstrap (BS) procedures or multivariate normal distributions (MVND). Here, we investigated the performance of an alternative method based on conditional distributions (CD). Covariates data from a hypertension drug development program were used. The methods were evaluated based on the original dataset (internal evaluation) and on their ability to reproduce an older, unobserved population (extrapolation). Similar results were obtained in the internal evaluation for summary statistics, yet BS was able to preserve the correlation structure of the empirical distribution, which was not adequately reproduced by MVND; CD was in between BS and MVND. BS does not allow to extrapolate to an unobserved population. When the dataset used to inform the extrapolation was well approximated by a MVND, the results from CD and MVND were comparable. However, improved extrapolation performance was observed for CD when deviations from normality assumptions occurred. If CTS is used to simulate within the observed distribution, BS is the preferred method. When extrapolating to new populations, a parametric method like CD/MVND is needed. In case the empirical multivariate distribution is characterized by linearly related covariates and unimodal marginal distributions, MVND can be used because of the simpler statistical framework and well-established use; however, if uncertainty about the MVND assumptions exists, CD will increase the confidence in the simulations compared to MVND.

Introduction

In several research fields that aim to study complex systems, mathematical modeling is routinely used to quantitatively describe the system, generate or test hypotheses and explore system behavior by means of model simulations. Drug discovery and development is one such field, that has seen the application of modeling and simulation - or pharmacometrics - growing at a fast pace in the last decades [1]. Arguably, the success of pharmacometrics lies in its ability to integrate and synthesize evidence collected from heterogenous sources throughout the drug development spectrum, ultimately providing a sound quantitative input to the decision-making process in pharmaceutical R&D [2-4].

In the pharmacometrician's toolkit, clinical trial simulation (CTS) offers the possibility to mitigate the risk of study failure by prospectively exploring the performance of a given study design and/or of the decision-making criteria [5]. Among others, applications of CTS include proof-of-concept study design [6], outcome prediction in phase 3 [7], pediatric trial optimization [8,9], evaluation of non-adherence [10,11] and guidance in dose titration decisions [12]. CTS can be broken down into three building blocks: a disease model, a drug model and a trial model [13]. The trial model requires the definition of a patient population through a set of characteristics, often termed covariates; in order to obtain realistic simulation scenarios, the set of covariates included in the CTS must be representative of the target population.

Common ways of generating covariate distributions for CTS are based upon non-parametric bootstrap (BS) procedures or multivariate normal distributions (MVND) [14,15]. While BS is a non-parametric method, MVND requires certain parametric assumptions to be met by the original data set at hand. The different nature of these two methods brings respective pros and cons: with BS it is

not possible to extrapolate outside of the observed distribution, but it preserves the relationships between covariates, thereby returning simulated subjects that are physiologically plausible. Conversely, the parametric distribution underlying the MVND - under certain assumptions about the means and correlations for the simulated population - can be used to generate completely new individuals. However, even with appropriate assumptions, the unrestricted nature of the underlying multivariate distribution may still lead to unrealistic simulated subjects.

Multiple imputation [16,17] is a technique used to handle complex missing data problems in statistical analysis, where multiple imputed data sets are analyzed separately, and inference is made on the pooled results. Missing data are imputed by sampling from conditional distributions using a pre-specified imputation model [18]. A commonly used imputation model is predictive mean matching (PMM) [19]. For each missing entry, PMM forms a small set of candidate donors from all complete cases that have predicted values closest to the predicted value for the missing entry. One donor is randomly drawn from the candidates, and the observed value of the donor is taken to replace the missing value. It follows that missing data are imputed with real values observed elsewhere in the data set, so imputations outside the observed data range will not occur. This feature makes sampling from conditional distributions using PMM (CD) an appealing method for covariate simulations, since in principle it embraces the advantages of MVND and BS, i.e., it can retain physiological plausibility while allowing to extrapolate outside of the observed multivariate distribution.

The objectives of this analysis were to investigate the operating characteristics of CD when used to simulate covariates distributions, and to compare them with those of BS and MVND.

Methods

Data

The data set used to build the simulations, hereafter referred to as the original data set, was obtained from a hypertension drug development program with data in 233 healthy subjects (HS) and 706 patients. The original data set contained the baseline values of the following covariates: age, weight (WT), serum creatinine (SCR), creatinine clearance (CRCL), sex and race. The summary statistics of the covariates in the original data set are reported in Table 1. The histograms of the continuous covariates stratified by HS vs. patients and the scatterplot matrix of the log-transformed continuous covariates are shown in Figure 1A and 1B, respectively.

Covariate simulation methods

BS. The BS method consisted of sampling, with replacement, covariate vectors from the original data set.

MVND. The MVND was implemented based on the framework presented in Tannenbaum et al. [15], that is briefly described. Before estimating the parameters of the MVND from the original data set, continuous covariates are log-transformed to constrain them to be positive; the covariate vectors simulated with the MVND are then exponentiated back to obtain the actual covariate values. As to categorical covariates, no pre-transformation is used, and they are treated as if they were continuous variables. This implies that the MVND can generate non-discrete values for categorical covariates, which then must be mapped back to their respective category based on a continuous critical value (see [15] for more details). Two important assumptions of the MVND are that the covariates in the original data set follow the same known distribution, and that they are linearly related to each other. A third assumption is that all the marginal distributions should not display clear multimodal trends. No truncation was applied to the covariates simulated with the MVND.

CD. Simulations with CD were performed using the fully conditional specification (FCS) algorithm (also known as multivariate imputation by chained equations) as implemented in the R package *mice* version 3.6.0 [19]. FCS is an iterative method in which missing data are imputed on a variable-by-variable basis. Let the covariate data set be represented by the matrix Y, FCS specifies a group of conditional densities $P(Y_j^{mis} \mid Y_j^{obs}, Y_{-j}, \square)$, where Y_j^{mis} and Y_j^{obs} are the set of missing and observed values for the j-th covariate, respectively, Y_{-j} are the values of all covariates in Y that are not the j-th one, and $\square$ is the imputation model. Starting from simple random draws from the marginal distribution, imputation under FCS is done by iterating over the conditionally specified imputation models (an illustrative example using standard linear regression as imputation model is shown in Figure S1). The complete FCS algorithm as implemented in the *mice* package is reported in Supplementary Information 1, further details can be found in [16,18].

The *mice* package offers the opportunity to choose from a selection of linear imputation models; here, the default methods were used, namely PMM [19] and multinomial logistic regression for continuous and categorical covariates, respectively. PMM first calculates the predicted value of the target covariate according to the specified linear regression model, secondly forms a set of donors/neighbors from the original data set constituted by the K closest values to the predicted one (K=5 in the present analysis), and finally makes the actual prediction by randomly drawing one value form the set of donors. Thus, the value of the simulated covariate is ultimately drawn from the original data set.

The CD simulations were performed as imputation of data sets where all covariates were assumed to be missing, leveraging the original data to train each univariate imputation model (Figure S1). By doing so, the probability of a certain covariate being missing does not depend on the unobserved values of that (or other) covariate(s), that is, the missing covariates are said to be missing at random

(MAR); this in turn implies that the missing data mechanism is ignorable, a condition that can be handled by the FCS algorithm [16,18]. An ad-hoc R function was developed and used to automate covariates simulation with CD (Supplementary Information 2).

Simulations set up and methods comparison

N=30 replicates of the same size as the original data set were simulated using BS, MVND and CD. CRCL was not directly simulated but derived from the simulated values of WT, SCR and sex using the Cockcroft-Gault [20] formula, similarly to what was done in the original data set.

In a first step, mean, standard deviation (SD), median, range and variance-covariance matrix for continuous covariates and proportions for categorical covariates were calculated for each of the 30 simulated replicates. Secondly, for each replicate *r*, the relative prediction error between the statistic of the simulated data set ($P_{sim,r}$) and the original statistic (P_{org}) was computed as rPE_r (%) = ($P_{sim,r}$ - P_{org})/P_{org}*100. Finally, relative bias (rBias) and root mean squared error (rRMSE) were derived as $\frac{1}{N}\sum_{r=1}^{N} rPE_r$ and $\sqrt{\frac{1}{N}\sum_{r=1}^{N}(rPE_r)^2}$, respectively.

The comparisons described above were carried out under three different scenarios. Scenario A used the entire original data set to inform the BS, MVND and CD simulations (internal evaluation). In Scenario B, with the aim of investigating the extrapolation performance of the methods, the original data set was divided into a training and a test data set, where only the training data set was used to inform the simulations, while the test data set was used to assess the predictive performance of the methods (external evaluation, P_{org} in this case is the statistics computed in the test data set). The training data set was defined as all subjects in the original data set younger than 55 years of age, mimicking a CTS exercise executed to investigate the features of a clinical trial performed in an

older, unobserved population. Scenario C was the same as Scenario B but excluding HS from the training data set (the histograms of the continuous covariates stratified by sex in the training and test data sets for Scenarios B and C are displayed in Figure S2). By definition, the BS does not allow to extrapolate outside of the empirical distribution, hence Scenarios B and C were tested for CD and MVND only. In Scenario B and C, the MVND simulations for age were obtained by truncating the distribution in the range 55-77 years. The age values simulated by the MVND were then used to seed the CD simulations, that is, the age values were not imputed but assumed to be known (see example C in Supplementary Information 2). Accordingly, only the covariance terms for age were included in the evaluation, whereas the other age statistics were not. It should be noted that in the CD simulations under Scenarios B and C the MAR assumption still holds true, as the probability of the missing data depends on the observed data, i.e., the probability of a covariate being missing is higher in older subjects compared to younger ones.

Results

Scenario A

Accuracy (rBias) and precision (rRMSE) of the continuous statistics are shown in Figure 2A-B. The mean and median of the original data set were maintained in the simulated population for all the methods. Likewise, SD and range were well estimated by BS and CD, whereas MVND resulted in larger rBias and rRMSE for the range of all covariates and for the SD of age. Figure 2C shows that the overall operating characteristics for the simulation of categorical covariates were similar across the three methods. BS was able to preserve the correlation structure of the empirical distribution, which was instead not adequately reproduced by MVND; CD was in between BS and MVND (Figure 3A). This in turn resulted in the lowest (BS) and highest (MVND) rBias and rRMSE in the elements of the variance-covariance matrix (Figure 3B-C).

Scenario B

Using CD for the simulation of an older patient population led to lower rBias and rRMSE in the summary statistics of the continuous covariates, compared to MVND (Figure 4A-B). Figure 4C shows that the prediction of the proportion of males and females in the extrapolated data set was more accurate and precise for MVND vs. CD, while similar performance was obtained for race.

Scenario C

When the HS data were removed from the training data set, no clear differences were observed between CD and MVND: some continuous covariates were better predicted by MVND (e.g. serum creatinine) and some others by CD (e.g. WT) (Figure 5A-B). As displayed in Figure 5C, sex was still more accurate and precise for MVND vs. CD.

Discussion

To maximize the predictability of a CTS exercise, plausible covariate values - representative of the target patient population - must be included in the trial model. Previous studies have investigated the operating characteristics of MVND as a tool to generate virtual populations [15], and compared them with BS approaches [14]. In the present work, we introduced an alternative approach to covariate simulations that borrowed the methodology used for multiple imputation of incomplete data sets in statistical analysis [16,17]. In addition, we have also explored the extrapolation performance of the methods in predicting the summary statistics of an older, unobserved patient population.

In the internal evaluation (Scenario A) the three methods adequately reproduced the summary statistics of the continuous covariates present in the original data set (Figure 2A-B). The results indicated that, for some covariates like age and SCR, dispersion parameters were not well predicted by MVND, which in turn led to the generation of implausible virtual subjects (e.g. 150 years old). Truncating the simulated covariates to more realistic values could help mitigating this problem; however, this was not assessed in the present analysis primarily because the cut-off values should be selected on a case-by-case basis, whereas the present work aimed at maximizing the generalizability of the results. In terms of categorical covariates, no clear differences between the operating characteristics of the three methods stood out; poorly represented categories, like “Other” and “American Indian” for race, were characterized by large rBias and rRMSE (Figure 2C).

Because the relationship between some pairs of log-transformed covariates deviated significantly from linearity (see e.g. age~WT correlation in Figure 1B), MVND provided more inaccurate and imprecise predictions of the covariance terms compared to BS and CD (Figure 3B-C). As expected, the BS could reproduce the original relationship between covariates regardless of its shape (Figure

3A), while CD appeared to be more robust than MVND to deviations from the linearity assumptions. It should be noted that the original data set was characterized by a substantial overlap in the covariate distributions between HS and patients (Figure 1); in case relevant differences between these distributions are observed, BS and MVND should account for that by using stratification and mixture modeling, respectively.

In Scenario B, the aim was to evaluate the extrapolation capabilities of MVND and CD by assessing their predictive performance on a test data set composed by subjects older than 55 years of age, when the simulations were informed by a training data set from younger subjects. Similar to Scenario A, the results suggested that in general CD achieved higher accuracy and precision levels than MVND for continuous covariates (Figure 4A-B). While comparable performance was observed for the prediction of the race categories, the predicted proportion of males and females was more biased and imprecise for CD compared to MVND. However, a previous study comparing MVND and CD approaches for multiple imputations found that CD is more accurate than MVND whenever the data include categorical variables [21]. Figure S2 indicates that - in the training data set - females have a generally lower WT than males, whereas in the test data set the WT distribution in male subjects is shifted towards that of females, and the two overlap into a unimodal distribution. It is reasonable to assume that, since WT is a relatively strong predictor for sex in the training data set, CD predicted a higher proportion of females (and lower proportion of males) because the mean WT in the test data set is lower than the one in the training data set. Instead, MVND led to better performance because the sex proportion is similar between the two data sets, though it should be noted that MVND can poorly perform when the marginal distributions are characterized by strong bimodalities [15].

When the HS data were removed from the training data set (Scenario C), the non-linearities in the correlation between age and other covariates were diminished (purple line in Figure 1B). As a result, the extrapolation performance of MVND improved compared to Scenario B and, although the CD operating characteristics also did, no large differences were observed between the two methods in terms of continuous covariates (Figure 4A-B). rBias and rRMSE in the sex covariate were still lower for MVND vs. CD, yet the aforementioned discussion on Scenario B concerning the different WT distribution between males and females applies to Scenario C as well (Figure S2).

In general, CD was more robust to departures of the data from normality assumptions. The reason seems to lie more on the PMM method used to predict the new covariate value rather than in the different statistical background between the two methods (in MVND the data are modeled as a sample from a joint multivariate normal distribution whereas in CD each variable is modeled conditionally on all the others). In fact, PMM slightly downweighs the model prediction by picking up a covariate value in the original data set that is very close to the predicted one; nevertheless, it should be noted that if strong non-linear relations exist in the original data set, the consequent misspecification in the underlying predictive model could result in poor CD performance [22]. Tannenbaum et al [15] suggest to carry out additional methods prior to creating the MVND in order to correct for non-linear relations between continuous covariates, but this was not performed in the present implementation (apart from log-transformation). Although this could have improved the MVND results, especially in terms of the variance-covariance matrix, it is likely that CD would have also benefitted from such pre-transformations given that it is still based on linear prediction models. Likewise, truncation of MVND simulations to plausible values might also have positively impacted the MVND operating characteristics. On the other hand, it can be noticed that while CD simulates ready-to-use covariates (Supplementary Information 2), MVND can require additional post-

processing of its input as well as output data, which in turn requires extra efforts and assumptions to be made.

The different results of the MVND between Scenario B and C can be attributed primarily to the distribution of age in HS, which is approximately uniform between 18 and 45 years. As shown in Figure 1A, the HS data distort the overall distribution of age, which is instead well approximated by a normal distribution in patients alone. Presumably, this is also the reason for the MVND issues in reproducing dispersion metrics under Scenario A. This suggests that study-specific nuisance covariates, such as HS vs. patients, should be handled carefully before embarking in a CTS effort. For example, in order to mimic the data generation process, the MVND approach could have been applied separately for patients and HS; however, this stratification should be done carefully as it may decrease the precision in the means and covariances of the two MVNDs, thereby increasing the uncertainty in the corresponding predictions.

BS was excluded from Scenario B and C because, by definition, it does not allow to extrapolate outside of the empirical distribution. This limitation can be overcome if observational data on the target indication are available, for example from databases based on routine clinical care used to identify patients for enrollment in clinical trials. However, while for common indications like hypertension this might easily be achieved, large patient registries for less frequent diseases are often not available.

CD performed with the *mice* package requires the specification of some options such as the number of donors K or the type of prediction model to use. All the defaults options were used in this analysis, yet the sensitivity of the results to the options choice should be further investigated. An interesting feature of *mice* that was not explored in the present work is that it potentially enables

simulation of time-varying covariates, which cannot be done with an MVND approach. Furthermore, extrapolation to other clinical contexts and generalizability of the results obtained in this work should be further explored.

The simulation of covariates, together with a disease and a drug model, represent the minimum working components that enable the generation of an in-silico response to treatment for a group of patients. Thus, besides providing a tool for prospective investigations of study design features, CTS represents an appealing methodology for the creation of synthetic controls in clinical trials. Synthetic controls are formally defined as statistical methods that can be used to evaluate the comparative effectiveness of an intervention using external control data [23]. Although synthetic controls are usually based on collected real evidence [24], the use of predictive statistical models to simulate virtual responses to standard-of-care treatments has recently been proposed as a way to generate synthetic controls [25]. In the case of rare diseases and small populations in general (e.g. pediatrics), where the use of control groups is often hampered by feasibility and ethical hurdles, simulated synthetic controls represent a promising alternative, as also indicated by their increased regulatory acceptance [26].

To conclude, the present analysis revealed that if CTS is used to simulate within the range of the observed distribution - for example when studies in the target population are already available from other sources - BS can be considered as the preferred method for covariates simulation, particularly because it is able to guarantee the physiological plausibility of the simulated covariates. On the other hand, if the CTS objectives involve extrapolation to new populations, a parametric method like CD or MVND is needed. In case the empirical multivariate distribution used to inform the extrapolation is characterized by linearly related covariates and unimodal marginal distributions, CD and MVND have comparable performance, and the use of MVND may be favored in light of its simpler

statistical framework and well-established use. However, if uncertainty about the MVND hypotheses exists, CD allows to partly relax these hypotheses, thereby increasing the confidence in the simulation outcomes compared to MVND.

Study Highlights

What is the current knowledge on the topic?

Common ways of generating virtual subjects to be used in clinical trial simulation (CTS) tasks are based upon bootstrap (BS) procedures or multivariate normal distributions (MVND).

What question did this study address?

What is the performance of an alternative method based on conditional distributions (CD) compared to BS and MVND?

What does this study add to our knowledge?

If CTS is used to simulate within the range of the observed distribution the BS is the preferred method for covariates simulation. When the CTS objectives involve extrapolations to new populations, a parametric method like CD or MVND is needed. As the MVND approach rests on relatively strong assumptions, i.e. linearly related covariates and unimodal distributions, CD is more robust when deviations from these assumptions occur.

How might this change drug discovery, development, and/or therapeutics?

Pharmacometricians will have at their disposal a new method for covariates simulation, which is particularly favorable when the empirical covariate distribution cannot be approximated by a MVND.

Acknowledgments

We would like to thank all Pharmetheus' colleagues for their valuable feedback on this work. Particular thanks go to Ola Caster, Sofia Friberg Hietala and Peter Milligan, whose critical input improved the quality of the manuscript.

Author contributions

G.S. and E.N.J. wrote the manuscript; E.N.J. designed the research; G.S. and E.N.J. performed the research; G.S. analyzed the data.

References

[1] EFPIA MID3 Workgroup, Marshall, S. et al. Good Practices in Model-Informed Drug Discovery and Development: Practice, Application, and Documentation. *CPT Pharmacometrics Syst. Pharmacol*. 5(3), 93-122 (2016).

[2] Milligan, P.A., Brown, M.J., Marchant, B., et al. Model-based drug development: a rational approach to efficiently accelerate drug development. *Clin. Pharmacol. Ther*. 93(6), 502-514 (2013).

[3] Gupta, N. et al. Transforming Translation Through Quantitative Pharmacology for High-Impact Decision Making in Drug Discovery and Development. *Clin. Pharmacol. Ther*. 107(6), 1285-1289 (2020).

[4] Marshall, S. et al. Model-Informed Drug Discovery and Development: Current Industry Good Practice and Regulatory Expectations and Future Perspectives. *CPT Pharmacometrics Syst. Pharmacol*. 8(2), 87-96 (2019).

[5] Holford, N., Ma, S.C. & Ploeger, B.A. Clinical trial simulation: a review. *Clin. Pharmacol. Ther*. 88(2), 166-182 (2010).

[6] Lockwood, P., Ewy, W., Hermann, D. & Holford, N. Application of clinical trial simulation to compare proof-of-concept study designs for drugs with a slow onset of effect; an example in Alzheimer's disease. *Pharm. Res*. 23(9), 2050-2059 (2006).

[7] De Ridder F. Predicting the outcome of phase III trials using phase II data: a case study of clinical trial simulation in late stage drug development. *Basic. Clin. Pharmacol. Toxicol*. 96(3), 235-241 (2005).

[8] Smania, G., Baiardi, P., Ceci, A., Magni, P. & Cella, M. Model-Based Assessment of Alternative Study Designs in Pediatric Trials. Part I: Frequentist Approaches. *CPT Pharmacometrics Syst. Pharmacol*. 5(6), 305-312 (2016).

[9] Smania, G., Baiardi, P., Ceci, A., Cella, M. & Magni, P. Model-Based Assessment of Alternative Study Designs in Pediatric Trials. Part II: Bayesian Approaches. *CPT Pharmacometrics Syst. Pharmacol*. 5(8), 402-410 (2016).

[10] Pérez-Ruixo, C., Remmerie, B., Peréz-Ruixo, J.J. & Vermeulen, A. A Receiver Operating Characteristic Framework for Non-adherence Detection Using Drug Concentration Thresholds-Application to Simulated Risperidone Data in Schizophrenic Patients. *AAPS J.* 21(3), 40 (2019).

[11] Piana, C., Danhof, M. & Della Pasqua, O. Impact of disease, drug and patient adherence on the effectiveness of antiviral therapy in pediatric HIV. *Expert Opin. Drug. Metab. Toxicol*. 13(5), 497-511 (2017).

[12] Clements, J.D., Perez Ruixo, J.J., Gibbs, J.P., Doshi, S., Perez Ruixo, C. & Melhem, M. Receiver Operating Characteristic Analysis and Clinical Trial Simulation to Inform Dose Titration Decisions. *CPT Pharmacometrics Syst. Pharmacol*. 7(11), 771-779 (2018).

[13] Gobburu, J.V. & Lesko, L.J.. Quantitative disease, drug, and trial models. *Annu. Rev. Pharmacol. Toxicol*. 49, 291-301 (2009).

[14] Teutonico, D. et al. Generating Virtual Patients by Multivariate and Discrete Re-Sampling Techniques. *Pharm. Res.* 32(10), 3228-3237 (2015).

[15] Tannenbaum, S.J., Holford, N.H., Lee, H., Peck, C.C. & Mould, D.R. Simulation of Correlated Continuous and Categorical Variables using a Single Multivariate Distribution. *J. Pharmacokinet. Pharmacodyn.* 33, 773–794 (2006).

[16] Van Buuren, S. *Flexible imputation of missing data*. (CRC press, Boca Raton, 2018).

[17] Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*. (John Wiley & Sons, New York, 1987).

[18] van Buuren, S. & Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. *J. Stat. Soft*, 45(3), 1-67 (2011).

[19] Little R.J.A. Missing Data Adjustments in Large Surveys. *JBES*, 6, 287-301 (1988).

[20] Cockcroft, D.W. & Gault, M.H. Prediction of creatinine clearance from serum creatinine. *Nephron*. 16(1), 31-41 (1976).

[21] Kropko, J., B. Goodrich, A. Gelman, and J. Hill. 2014. Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches. *Polit Anal* 22(4), 497-519 (2014).

[22] Morris, T.P., White, I.R. & Royston, P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med. Res. Methodol.* 14, 75 (2014).

[23] Thorlund, K., Dron, L., Park, J.J.H. & Mills, E.J. Synthetic and External Controls in Clinical Trials - A Primer for Researchers. *Clin Epidemiol*. 12, 457-467 (2020).

[24] Gökbuget, N. et al. Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. *Blood Cancer. J.* 23;6(9):e473 (2016).

[25] Switchenko, J.M., Heeke, A.L., Pan, T.C., Read, W.L. The use of a predictive statistical model to make a virtual control arm for a clinical trial. *PLoS ONE*, 14(9) (2019).

[26] Enpr-EMA working group on trial preparedness. Preparedness of medicines' clinical trials in paediatrics. 2020 <https://www.ema.europa.eu/en/documents/other/preparedness-medicines-clinical-

trials-paediatrics-recommendations-enpr-ema-working-group-trial_en.pdf> (2020). Accessed 25 October 2020.

Figures

Figure 1. Histograms of continuous covariates colored by patients vs. healthy subjects (A) and correlation between pairs of continuous covariates (B). The solid line is a loess regressor.

Figure 2. Relative Bias and RMSE for summary statistics of continuous (A and B) and categorical (C) covariates in Scenario A (cells where the absolute value of relative Bias/RMSE is greater than 30% are grayed out).

Figure 3. Visual predictive check of the relationship between pairs of continuous covariates (A): 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. Relative Bias (B) and RMSE (C) for the variance-covariance matrix of the continuous covariates (cells where the absolute value of relative Bias/RMSE is greater than 30% are grayed out). Scenario A.

Figure 4. Relative Bias and RMSE for summary statistics of continuous (A) and categorical (C) covariates in Scenario B (cells where the absolute value of relative Bias/RMSE is greater than 30% are grayed out). Relative Bias and RMSE for the variance-covariance matrix of the continuous covariates (B) in Scenario B (cells where the absolute value of relative Bias/RMSE is greater than 100% are grayed out). There were no subjects with “race = other” in the training data set, therefore in subfigure C rBias and rRMSE were computed using the absolute difference instead of the relative one for this category.

Figure 5. Relative Bias and RMSE for summary statistics of continuous (A) and categorical (C) covariates in Scenario C (cells where the absolute value of relative Bias/RMSE is greater than 30% are grayed out). Relative Bias and RMSE for the variance-covariance matrix of the continuous covariates (B) in Scenario C (cells where the absolute value of relative Bias/RMSE is greater than 100% are grayed out).

Table 1. Summary statistics of the covariates in the original data set.

Supplementary information titles

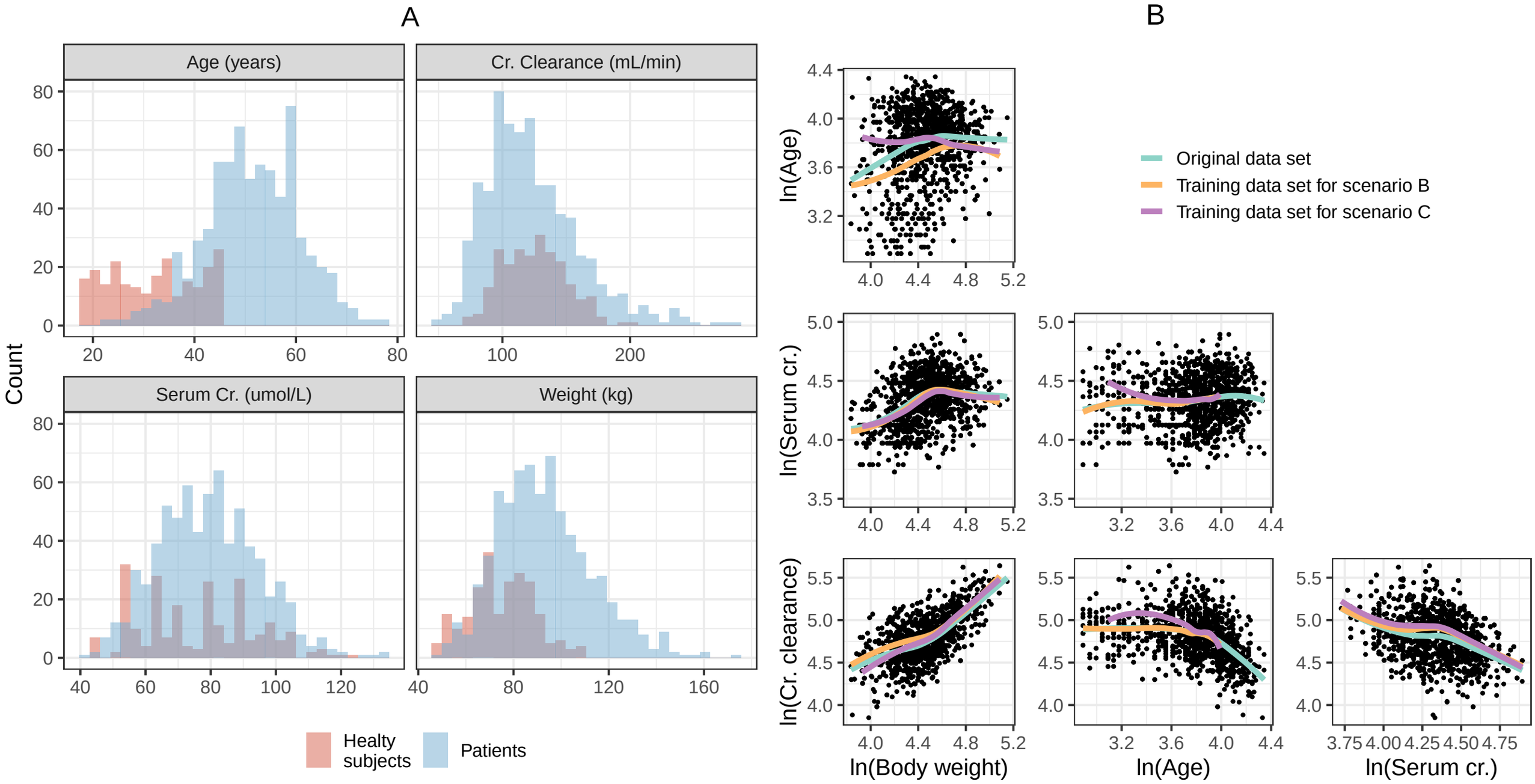
Supplementary Information 1: Fully conditional specification (FCS) algorithm as implemented in the *mice* package.

Supplementary Information 2: R function developed for covariates simulation using CD with minimum working examples of use.

Figure S1: Illustrative example of the methodology used to simulate covariates with CD.

Figure S2: Histograms of the continuous covariates in the training and test data sets for Scenarios B and C, colored by sex.

Total number of subjects	939
Continuous covariates	
Age (years)	
Mean	46.4
SD	12.4
Median	47.0
Range	18.0-77.0
Weight (kg)	
Mean	88.5
SD	20.1
Median	86.6
Range	46.3-172
Serum creatinine (µmol/L)	
Mean	78.5
SD	16.2
Median	78.7
Range	41.5-133
Creatinine clearance (mL/min)	
Mean	124
SD	34.3
Median	120
Range	47.0-282
Categorical covariates	
Population, n (%)	
Patients	706 (75.2)
Healthy subjects	233 (24.8)
Sex, n (%)	
Males	534 (56.9)
Females	405 (43.1)
Race, n (%)	
White	737 (78.5)
Black	179 (19.1)
Asian	19 (2.0)
American Indian	2 (0.2)
Other	2 (0.2)


A
Age (years)
Cr. Clearance (mL/min)
Serum Cr. (umol/L)
Weight (kg)
Count
Healty subjects
Patients
B
Original data set
Training data set for scenario B
Training data set for scenario C
ln(Age)
ln(Serum cr.)
ln(Cr. clearance)
ln(Body weight)
ln(Age)
ln(Serum cr.)

A

	BS				CD				MVND			
	AGE	CRCL	SCR	WT	AGE	CRCL	SCR	WT	AGE	CRCL	SCR	WT
SD	0.04	−1.36	0.02	−1.24	−0.39	2.91	−0.59	1.09	17.83	2.29	3.17	0.26
RANGE	−0.11	−3.09	−1.67	−5.48	−0.06	4.66	−2.82	−4.12	73.95	10.57	24.27	11.3
MEDIAN	0	0.03	0.12	−0.08	0.35	−1.03	−0.51	−0.4	−5.47	−0.04	−2.6	−0.2
MEAN	0.08	0.01	−0.19	−0.05	−0.12	0.32	−0.26	0.04	0.46	−0.42	−0.07	−0.06

Bias (%)

30
20
10
0
−10
−20
−30

B

	BS				CD				MVND			
	AGE	CRCL	SCR	WT	AGE	CRCL	SCR	WT	AGE	CRCL	SCR	WT
SD	1.89	3.22	2.01	3.44	3.01	5.89	2.32	2.59	18.19	3.45	4.15	2.61
RANGE	0.44	4.88	2.68	7.97	0.31	14.09	4.68	6.24	78.53	17.62	28.1	16.64
MEDIAN	1.35	0.95	0.96	0.83	1.6	2.05	1.29	1.25	5.59	1.03	2.75	0.84
MEAN	0.98	0.85	0.5	0.84	1.56	1.62	0.9	1.04	1.11	0.89	0.69	0.74

RMSE (%)

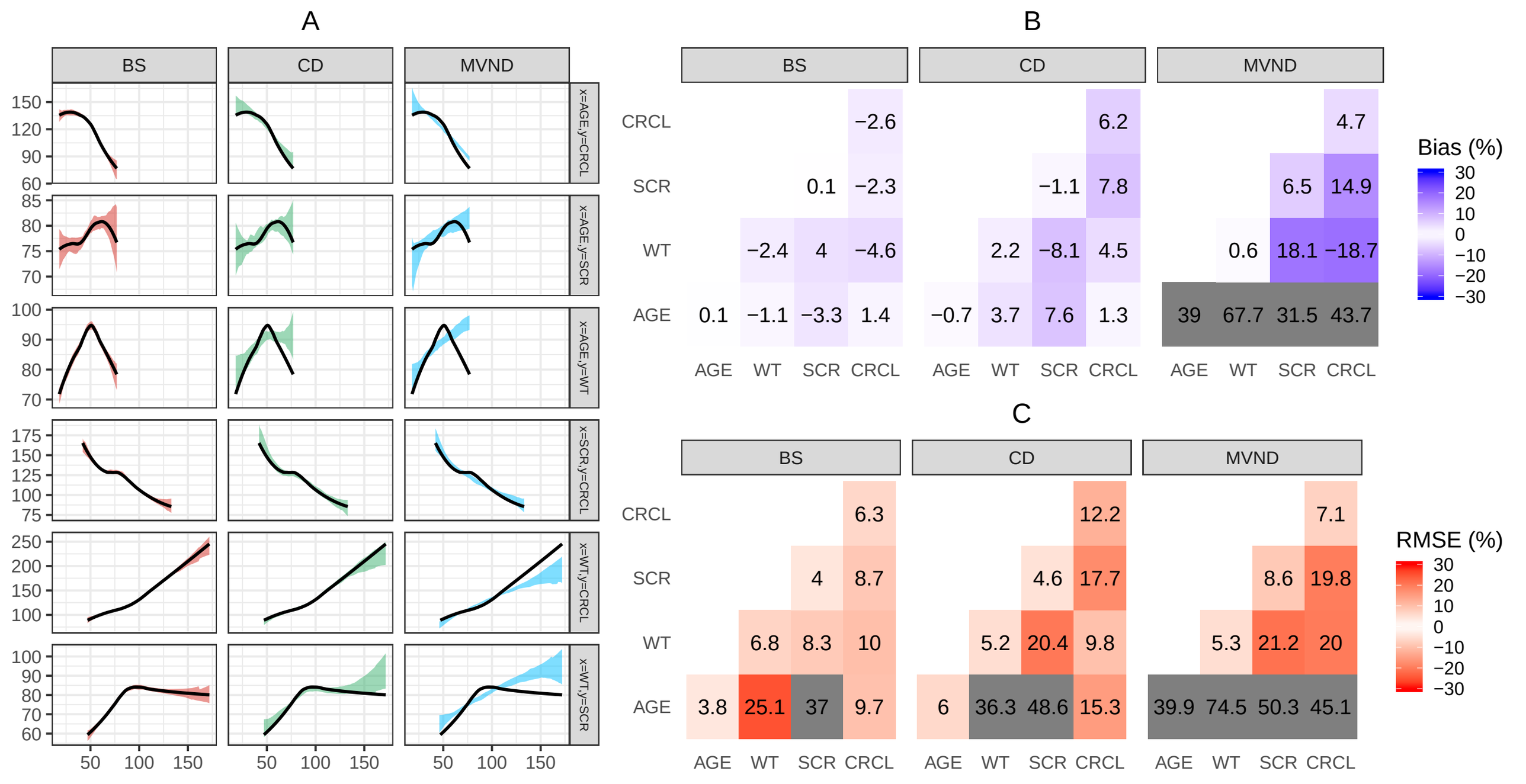
30
20
10
0

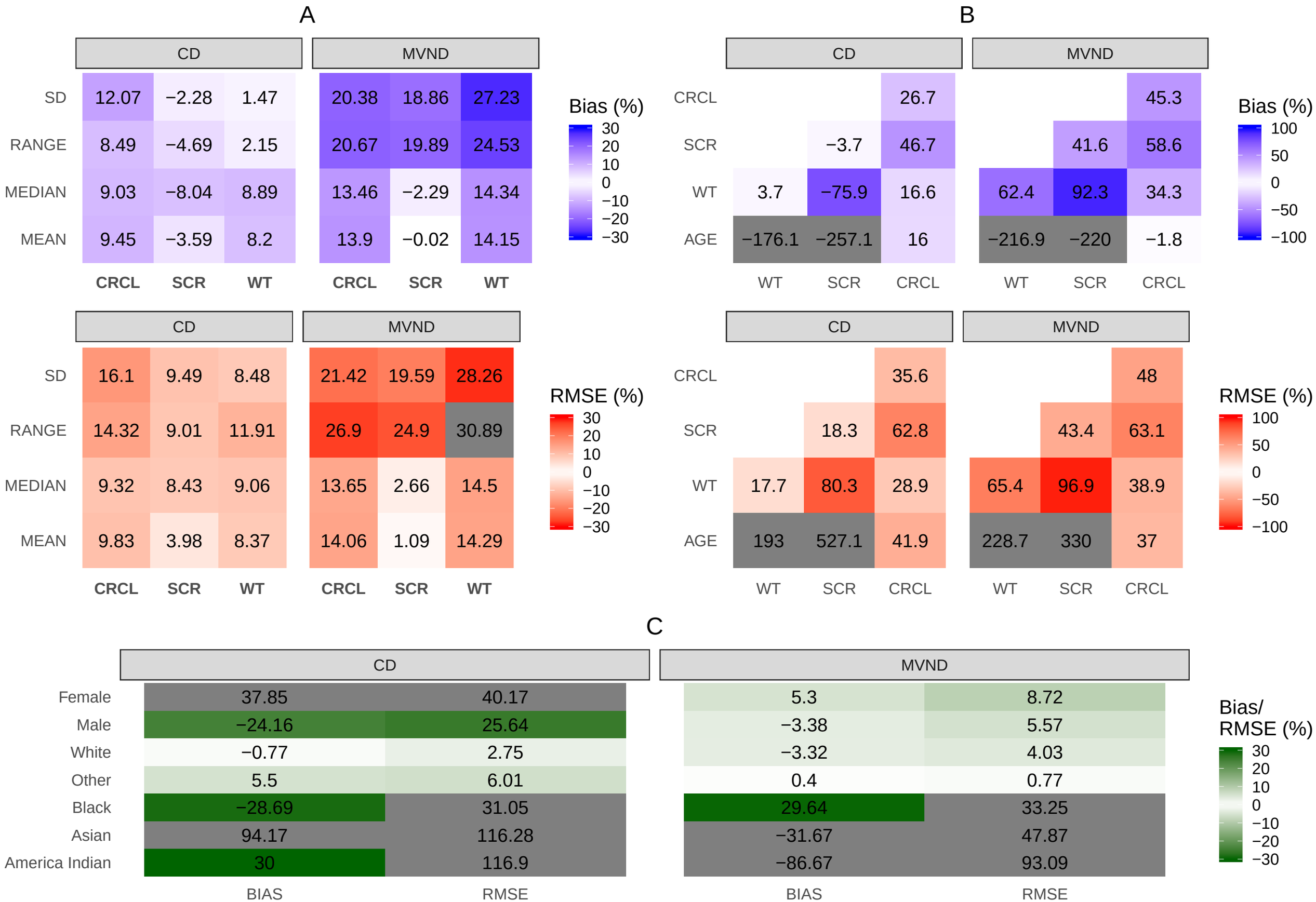
C

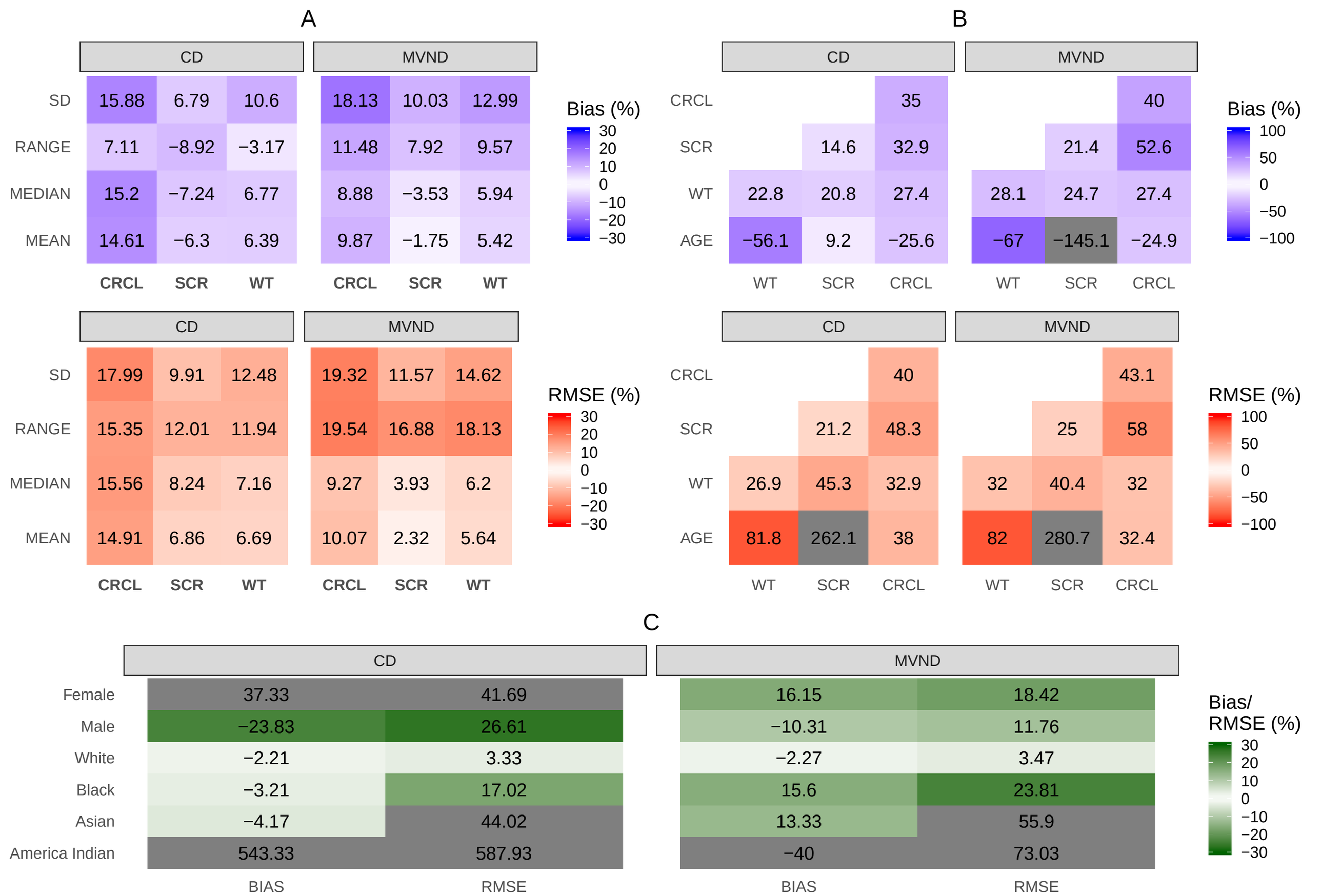
	BS		CD		MVND	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
Female	0.76	3.77	2.11	5.7	1.23	4.16
Male	−0.57	2.86	−1.6	4.33	−0.93	3.16
White	0.34	1.55	−0.93	1.9	−0.41	1.48
Other	25	79.06	76.67	121.11	13.33	74.16
Black	−1.47	6.81	2.1	6.12	1.01	5.6
Asian	−0.88	18.51	1.93	25.22	4.39	21.42
America Indian	−11.67	65.19	58.33	96.18	5	58.45

Bias/
RMSE (%)

30
20
10
0
−10
−20
−30


A
BS
CD
MVND
x=AGE,y=CRCL
x=AGE,y=SCR
x=AGE,y=WT
x=SCR,y=CRCL
x=WT,y=CRCL
x=WT,y=SCR
B
Bias (%)
CRCL −2.6 6.2 4.7
SCR 0.1 −2.3 −1.1 7.8 6.5 14.9
WT −2.4 4 −4.6 2.2 −8.1 4.5 0.6 18.1 −18.7
AGE 0.1 −1.1 −3.3 1.4 −0.7 3.7 7.6 1.3 39 67.7 31.5 43.7
AGE WT SCR CRCL
C
RMSE (%)
CRCL 6.3 12.2 7.1
SCR 4 8.7 4.6 17.7 8.6 19.8
WT 6.8 8.3 10 5.2 20.4 9.8 5.3 21.2 20
AGE 3.8 25.1 37 9.7 6 36.3 48.6 15.3 39.9 74.5 50.3 45.1
AGE WT SCR CRCL


A
CD
MVND
SD 12.07 −2.28 1.47 20.38 18.86 27.23
RANGE 8.49 −4.69 2.15 20.67 19.89 24.53
MEDIAN 9.03 −8.04 8.89 13.46 −2.29 14.34
MEAN 9.45 −3.59 8.2 13.9 −0.02 14.15
CRCL SCR WT CRCL SCR WT
Bias (%)
30 20 10 0 −10 −20 −30
CD
MVND
SD 16.1 9.49 8.48 21.42 19.59 28.26
RANGE 14.32 9.01 11.91 26.9 24.9 30.89
MEDIAN 9.32 8.43 9.06 13.65 2.66 14.5
MEAN 9.83 3.98 8.37 14.06 1.09 14.29
CRCL SCR WT CRCL SCR WT
RMSE (%)
30 20 10 0 −10 −20 −30
B
CD
MVND
CRCL 26.7 45.3
SCR −3.7 46.7 41.6 58.6
WT 3.7 −75.9 16.6 62.4 92.3 34.3
AGE −176.1 −257.1 16 −216.9 −220 −1.8
WT SCR CRCL WT SCR CRCL
Bias (%)
100 50 0 −50 −100
CD
MVND
CRCL 35.6 48
SCR 18.3 62.8 43.4 63.1
WT 17.7 80.3 28.9 65.4 96.9 38.9
AGE 193 527.1 41.9 228.7 330 37
WT SCR CRCL WT SCR CRCL
RMSE (%)
100 50 0 −50 −100
C
CD
MVND
Female 37.85 40.17 5.3 8.72
Male −24.16 25.64 −3.38 5.57
White −0.77 2.75 −3.32 4.03
Other 5.5 6.01 0.4 0.77
Black −28.69 31.05 29.64 33.25
Asian 94.17 116.28 −31.67 47.87
America Indian 30 116.9 −86.67 93.09
BIAS RMSE BIAS RMSE
Bias/ RMSE (%)
30 20 10 0 −10 −20 −30


A
CD
MVND
SD
RANGE
MEDIAN
MEAN
15.88 6.79 10.6 18.13 10.03 12.99
7.11 −8.92 −3.17 11.48 7.92 9.57
15.2 −7.24 6.77 8.88 −3.53 5.94
14.61 −6.3 6.39 9.87 −1.75 5.42
CRCL SCR WT CRCL SCR WT
Bias (%)
30 20 10 0 −10 −20 −30
17.99 9.91 12.48 19.32 11.57 14.62
15.35 12.01 11.94 19.54 16.88 18.13
15.56 8.24 7.16 9.27 3.93 6.2
14.91 6.86 6.69 10.07 2.32 5.64
RMSE (%)
30 20 10 0 −10 −20 −30
B
CD
MVND
CRCL
SCR
WT
AGE
35 40
14.6 32.9 21.4 52.6
22.8 20.8 27.4 28.1 24.7 27.4
−56.1 9.2 −25.6 −67 −145.1 −24.9
WT SCR CRCL WT SCR CRCL
Bias (%)
100 50 0 −50 −100
40 43.1
21.2 48.3 25 58
26.9 45.3 32.9 32 40.4 32
81.8 262.1 38 82 280.7 32.4
WT SCR CRCL WT SCR CRCL
RMSE (%)
100 50 0 −50 −100
C
CD
MVND
Female
Male
White
Black
Asian
America Indian
37.33 41.69 16.15 18.42
−23.83 26.61 −10.31 11.76
−2.21 3.33 −2.27 3.47
−3.21 17.02 15.6 23.81
−4.17 44.02 13.33 55.9
543.33 587.93 −40 73.03
BIAS RMSE BIAS RMSE
Bias/
RMSE (%)
30 20 10 0 −10 −20 −30