Research Master’s programme: Methodology and Statistics for the Behavioural, Biomedical and Social Sciences
Utrecht University

Research Report Anaïs Fopma (6199356)
Title: A blended distance to define "people-like-me"
Date: 09-05-2022

Supervisors: Prof. Dr. Stef van Buuren, Dr. Gerko Vink & Mingyang Cai

Preferred journal of publication: Statistics in Medicine
Word count: —-

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

MASTER'S THESIS

A blended distance to define "people-like-me"

Anaïs M. Fopma*

[1]Methodology & Statistics, Utrecht University, the Netherlands

Correspondence
*A.M. Fopma, Padualaan 14, 3584 CH Utrecht,the Netherlands. Email: a.m.fopma@uu.nl

Present Address
Padualaan14,3584CH Utrecht, the Netherlands

Summary

Curve matching is a powerful tool to predict the development of a child (the target) with the data of other children (donors). The technique relies on predictive mean matching, which matches donors that are most similar to the target based on the predictive distance. Even though this approach ensures high prediction accuracy, there are two disadvantages. Firstly, it requires users of curve matching to select a particular future time point to base the matches on. In some cases, it may be difficult to choose this time point. Secondly, the predictive distance may make matches look unconvincing, as the profiles of the matched donors can substantially differ from the profile of the target, even if they are close on the predicted time point. To counterbalance these disadvantages, similarity between the curves of the donors and the target can be taken into account when selecting donors. The objective of the current study is to do so by combining the predictive distance with the Mahalanobis distance, thus creating a 'blended distance' measure.

KEYWORDS:
Curve matching, predictive mean matching, distance measures, metrics

1 | INTRODUCTION

The first three years of childhood form a crucial stage in determining children's subsequent development and health outcomes.[1] For this reason, growth monitoring is considered to be an integral part of paediatrics. It can aid in the identification of problems in development such as growth stunting, and ensure timely treatment or intervention to improve the child's health.[2] However, growth monitoring solely provides insights in the past and current developmental stages of the child. Growth curve modeling, on the other hand, can be used to predict future development. It could therefore provide more specific answers to questions health professionals, parents, and insurance companies may have, such as: 'Given what I know of the child, how will it develop in the future?' and 'Does this child get the most effective treatment available?'[3]

1.1 | Curve matching

An approach currently used for growth curve modeling is curve matching. Curve matching[3] is a nearest neighbour technique for individual prediction that constructs a prediction by aggregating the histories of "people-like-me". Its aim is to predict the growth of a target child by using the data of other children that are most similar to the target child.

In order to select these donors, some form of similarity needs to be defined to match the donors to the target child. Therefore, the key question is: How are good matches obtained? The current approach uses predictive mean matching (PMM). PMM makes use of an existing donor database, containing the growth data of children who are older than the target child, and of which information at a later age is available. The first step is to fit a linear regression model on the donor database. Then, this model is used to predict the values for all donors and for the target at a certain point in the future, for example at 14 months. Finally, the distance between the predicted value of each of the donors and the predicted value of the target is calculated, which is referred to as the predictive distance. A number of donors – usually five - with the smallest predictive distance are selected as the best matches. Their growth curves are then plotted and point estimates can be calculated by averaging the measurements. The growth patterns of the matched children thus suggest how the target child might develop in the future.

An advantage of this technique is its high prediction accuracy.[3] Moreover, the applications of curve matching can be extended to settings other than the prediction of child development, such as patient recovery after an operation, prediction of longevity, and decision-making when multiple treatments are available.[3]

1.2 | Alternative approach

Even though PMM has proven to be promising in growth curve matching,[3] there are two reasons to move beyond the predictive distance used in PMM and investigate an alternative metric. Firstly, PMM requires users of curve matching to select a particular future time point to base the matches on (e.g. 14 months of age). In some cases, it may be difficult to choose this time point, especially when the 'future' is more vaguely defined as a time interval.[4] Secondly, the predictive distance may make the matches look unconvincing. The trajectories of the selected donors may all be close to the prediction for the target child at 14 months, but this does not imply that the histories are identical. After all, different profiles may lead to the same predicted value. Consequently, the curves of some of the matches may be quite far from the curve of the target child. Some users of curve matching feel that such discrepancies are undesirable, as these matches do not appear to be *people-like-me*.[4]

For these reasons, the practical implementation and use of curve matching can be improved by combining the predictive distance with the Mahalanobis distance, thus creating a "blended distance" measure. This blended metric would take into account historical similarity, by giving more weight to similarities between units in the full predictor space. The objective of this study is to investigate what the properties of such a blended distance measure are.

2 | METHODS

Two simulations studies will be conducted. The following sections describe each study in accordance with the ADEMP-structure for reporting simulation research.[5] The different versions of the blended metric (i.e. the methods), the aims of the study, the data-generating mechanisms, and the estimand and performance measures are discussed. In addition, an application of the blended metric to an empirical data set is outlined.

2.1 | Simulation study I

2.1.1 | Blended metric

The blended distance measure to be evaluated is a weighted version of the predictive distance and the Mahalanobis distance. As described above, the predictive distance is the distance between the predicted value of a donor and the predicted value of the target at a particular future time point. The Mahalanobis distance is defined as the distance between two N dimensional points scaled by the variation in each component of the point. For example, if $\vec{x}$ and $\vec{y}$ are two points from the same distribution which has covariance matrix $\mathbf{C}$, then the Mahalanobis distance is given by

$$((\vec{x}-\vec{y})'\mathbf{C}^{-1}(\vec{x}-\vec{y}))^{\frac{1}{2}}. \tag{1}$$

Two potential versions of the blended metric will be compared: one that uses ranking and one that uses scaling. The ranked blended metric is created as follows. First, the predictive distance and the Mahalanobis distance are calculated for each donor. Then, the k donors with a low value for both the predictive distance and the Mahalanobis distance are selected. In order to do so, the rank is calculated for the predictive distance PD and the Mahalanobis distance MD, where ties are randomly broken. The ranked blended distance RBD is then given by:

$$RBD = p \cdot \text{rank}_{PD} + (1-p) \cdot \text{rank}_{MD}, \tag{2}$$

where rank_{PD} is the rank for the predictive distance PD, rank_{MD} is the rank for the Mahalanobis distance MD, p is the blending factor (or weight) assigned to rank_{PD}, and $0 \leq p \leq 1$. The k donors with the lowest values on RBD are selected as the best matches.

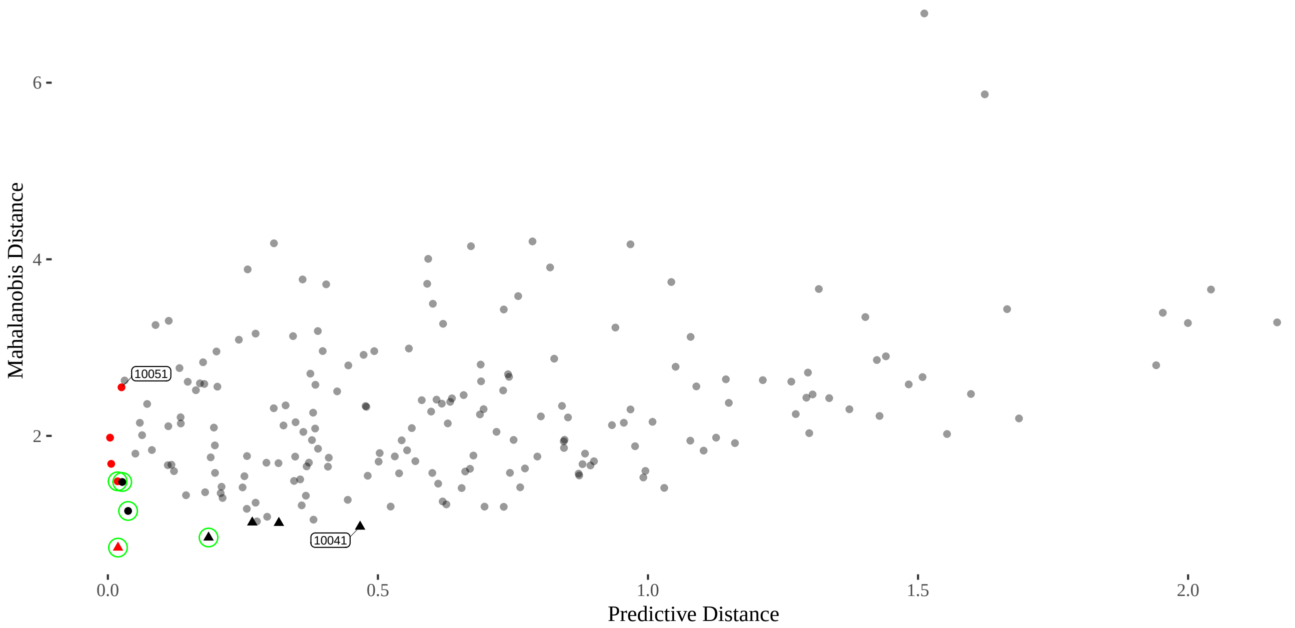

FIGURE 1 Mahalanobis distance plotted against predictive distance for each of the 199 donors. The donors in red are the five matches with the smallest predictive distance, the triangular donors those with the smallest Mahalanobis distance, and the donors circled in green those with the smallest blended distance.

The scaled blended metric is created similarly, but scales the predictive distance *PD* and the Mahalanobis distance *MD* before combining them. The scaled blended distance *SBD* is then given by:

$$SBD = p \cdot \frac{PD - \bar{x}_{PD}}{\sigma_{PD}} + (1 - p) \cdot \frac{MD - \bar{x}_{MD}}{\sigma_{MD}}, \tag{3}$$

where $\bar{x}_{PD}$ is the mean of the predictive distances, σ_{PD} their standard deviation, $\bar{x}_{MD}$ is the mean of the Mahalanobis distances, and σ_{MD} their standard deviation.

In theory, these two versions of the blended distance should yield identical results. However, the scaled version would be computationally more efficient. Both versions are included in this study to confirm that they produce the same results, and if this is indeed the case, the scaled version could be implemented in the `mice`[6] package.

As an example, the blended distance is illustrated in Figure 1. Here, the data of 200 children from the *Sociaal Medisch Onderzoek Consultatiebureau Kinderen* (SMOCK) study are used.[7] The first subject is taken as the target, the 199 other subjects as the donors. For all donors, the Mahalanobis distance for the measurements during the first six months of growth is calculated. In addition, the predictive distance between each donor and the target is calculated. In the figure, the Mahalanobis distance and predictive distance are plotted against each other. The red donors are the five matches with the smallest predictive distance, where especially subject 10051 has a large Mahalanobis distance. The triangular donors are the five matches with the smallest Mahalanobis distance, where especially subject 10041 has a large predictive distance. A weighted blended distance measure would balance the distance measures, such that the donors with a low value for both distance measures are chosen. These are circled in green.

In the current study, blending factors of respectively 1, 0.5, and 0 will be evaluated for the blended metric. A blending factor of 1 implies that the blended distance is equal to the predictive distance, whereas a weight of 0 implies that it is equal to the Mahalanobis distance. Therefore, a blending factor of 0.5 gives equal weight to both distance measures. PMM will be used as a reference in order to evaluate whether we do indeed obtain the same results with a blending factor of 1. Using PMM, both the ranking and scaling methods with three different blending factors each, this results in seven different versions of the blended metric to be evaluated.

2.1.2 | Aims

The main objective of this study is to investigate what the properties of the blended metric are. More specifically, we want to answer the following questions:

1. Do a ranked and a scaled version of the blended distance measure yield identical results?
2. Does a blending factor of 1 yield results identical to those obtained by PMM, as intended?
3. How is the performance of the blended metric related to the missingness mechanism, the proportion of missingness in the data, the distribution of the data, and the correlation in the data?
4. Is there a penalty from blending in terms of reduced predictability?

It is expected that the ranked and scaled versions do yield identical results, and that blending with a factor of 1 does indeed give the same results as PMM does. As pointed out before, PMM has been shown to result in high prediction accuracy. Therefore, it is expected that the predictability of the blended distance will decrease as the blending factor favours the Mahalanobis distance. When the correlation in the data is low, the prediction model will fit poorly and the blended metric is expected to perform worse when more weight is given to the predictive distance. When the correlation in the data is high, the prediction model will fit better, and the prediction model will explain more variance in the outcome. In this case, the blended metric is expected to perform better when more weight is given to the predictive distance. Finally, it is expected that the blended metric will perform worse in skewed data, when more weight is given to the Mahalanobis distance.

2.1.3 | Data-generating mechanisms

I don't know if this section is still correct after redesigning the simulations? In order to answer the previous questions, the blended distance measure will be evaluated in simulated data that meet different conditions. All data are generated from one of 24 data-generating mechanisms, with equal means, but with varying missingness proportions, missingness mechanisms, distributions, and variance-covariance matrices.

Three continuous predictor variables $\mathbf{X_1}$, $\mathbf{X_2}$, and $\mathbf{X_3}$ are defined, corresponding to the standardised height measurements (Z-scores) at birth, at 1 month, and at 2 months. One continuous outcome $\mathbf{Y}$ is defined, corresponding to the predicted standardised height measurement at the age of 14 months.

The distribution of the data is varied over two conditions. The data generating mechanism of the predictor space is a multivariate normal distribution for the first condition, and a strongly skewed multivariate distribution for the second condition, $\mathbf{X} = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, with mean vector $\boldsymbol{\mu} = [0, 0, ..., 0]$. In order to achieve this, the predictors are transformed,[8] where

$$X = X$$

for the first condition and

$$X = X^{12}/max\{X^{11}\}$$

for the second condition.

The correlation in the data is varied over three conditions. The covariance matrix Σ for the populations with two predictors is given by:

$$\Sigma = \begin{bmatrix} 1 & \sigma^2\rho & \sigma^2\rho \\ & 1 & \sigma^2\rho \\ & & 1 \end{bmatrix},$$

where the off-diagonal elements are 0 for the first condition, 0.1 for the second condition, and 0.7 for the third condition.

The proportion of missingess in the outcome variable is varied over two conditions. The first condition simulates a setting with 25% missingness, the second a setting with 50% missingness.

Finally, the missingness mechanism is varied over two conditions. The first concerns a missing completely at random (MCAR) mechanism, where missingness does not depend on the values of the data, missing or observed.[9] The second concerns a missing at random (MAR) right mechanism. This means that missingness does depend on the data, but only through observed components of the data.[9] Is there a reference I can use to explain that the latter is the more extreme mechanism?

We consider a full-factorial simulation study design, where each of the possible combinations of weighting and data-generating mechanisms are evaluated. As there are seven different methods to be evaluated and 24 different data-generating mechanisms, the simulation will yield 168 results. From each data-generating mechanism, a sample of size 500 is drawn. The number of simulations run for each setting is set to 1000.

Note to self: explain that the simulation is seed-dependent, and add the same simulation with a different seed to the research archive like Hanne said

2.2 | Estimand and performance measures

The estimands of interest in this study are the predicted or imputed values. To assess the performance of each metric under each combination of conditions, the parameter estimate (qbar), standard error (se), total variance about the parameter estimate (t), degrees of freedom (df), variance between imputations (b), upper and lower 95% confidence limits, true value, coverage, and bias are computed.

2.3 | Simulation study II

Notes:

1. Now you remove a single value out of 500 and impute it »m times (lets say 50). Evaluate it against its original value.
2. Make a subset in the 500 cases (let's say 25 cases) that together have similar predictive distances, but are dissimilar in their trajectories. and then you do (1), but you make sure that only one out of the 25 is removed.

2.3.1 | Aims

The objective of the second simulation study is to evaluate the blended metric that performed best in Simulation study I. Or do we do all of them? If the results show that there is no clear distinction between which performed best, the blended metric with a blending factor of 0.5 will be evaluated. In this second study, a setting will be simulated that reflects the practical implementation of the blended metric in the prediction of child growth. The questions of interest for this study are:

1. How does the blended metric perform in data that resembles a practical context?
2. Does the blended metric perform better when the donors are dissimilar in their trajectories?

I was wondering if we could not use the empirical data for this purpose, instead of doing another simulation study. Because, the empirical data already are data from practice, and I am also not sure what to do with the empirical data. Would I just test each of the blended metrics on these data once, and see if the predictions come close to the actual height of the target child? Would it make sense to do this?

2.3.2 | Data-generating mechanisms

Data is simulated from a single data-generating mechanism. The same variables are defined as in the first simulation study, where the data is normally distributed and the off-diagonal elements in the covariance are set to 0.7. A sample of size 500 is drawn. To simulate the prediction of the height measurement at 14 months for a single target child, the missingness proportion is set to 0.02. The simulation study consists of two steps. The first is to predict the target's height by means of the full donor set of 499 donors. The second is to predict the target's height by means of a subset of 24 donors, of which the predictive distance is similar, but of which the growth trajectories are dissimilar. But is this not the same as using a blended distance with a low blending factor? A 100 simulations are run for each step.

2.3.3 | Estimand and performance measures

The estimand is the single value to be imputed, which in practice would be equivalent to the predicted height measurement of the target at 14 months of age. The performance measure of interest is the bias between the predicted value and the original value.

2.4 | Study on empirical data

After the simulation study is conducted, the blended metric will be evaluated in an application to empirical data from the SMOCK study.[7] The weights used will be the same as to those used in the simulation study: 1, 0.5, and 0. The SMOCK database contains the anonymised growth data of 1,933 children aged 0-15 months. In addition, the database contains covariates that influence growth, such as the sex, gestational age, birth weight, and height of the father and mother.

The growth data in the SMOCK database consist of the height measurements of children at different observation times. It is important to note that the actual time points of data collection will sometimes differ substantially from the scheduled times. This may be due to a doctor's visit being planned during a holiday, the subject not showing up at the appointment, or the measurement device being out of order at the time of the scheduled observation.[4] As a consequence, the observation times will vary across subjects, and are said to be irregular. Irregular observation times present significant challenges for quantitative analysis, as it becomes more complex to predict the future from past data. Usually, a linear mixed model with time-varying outcomes is applied. However, an alternative is offered by the broken stick model,[4] which converts irregularly observed data into a set of repeated measures. As a result, each child's growth trajectory can be approximated by a series of connected straight lines. The breakpoints between these lines are set to be the pre-specified, scheduled observation times. The advantage is that repeated measures data offer a lot more simplicity than the use of linear mixed models. Therefore, the empirical data will be analysed using the broken stick model.

2.5 | Software

R version 4.1.2 (2021-11-01)[10] will be used to simulate the data and perform the analyses. The `mice.impute.pmm` function in the `mice`[6] package will be used to perform PMM and an adaptation of this function will be used to calculate the blended distance. As the empirical data consist of irregular observation times, the `brokenstick` package[4] will be used for estimating the growth models. Instructions and scripts to reproduce the simulation results are available in the research archive of this project.

3 | RESULTS

3.1 | Simulation study I

The simulation results for each of the seven methods are displayed in Table A1 through Table A7 in Appendix A. In each table, the data-generating mechanisms are specified in the left columns, by indicating the missingness mechanism, missingness proportion, skewness of the distribution, and correlation in the data. The results for the coverage are visualised in Figure 2 and those for the bias are visualised in Figure 3. The results are discussed below on the basis of the research questions.

3.1.1 | Comparison of ranked and scaled blended metric

It was expected that the ranked and scaled versions of the metrics would yield identical results. The results show that this is true in some cases, but that not all are identical. When comparing the results for blending factor = 1 in Table A2 and Table A5, they show that the scaled method yields slightly higher coverages but larger biases overall. When comparing the results for blending factor = 0.5 in Table A3 and Table A6, they show that the scaled method outperforms the ranked method. The results for blending factor = 0 in Table A4 and Table A7 show that the ranked method outperforms the scaled method. Overall, the ranked method performs slightly better, except when the blending factor is set to 0.5. As most of the results are similar, however, the scaled version of the blended metric might be preferable, as it is computationally more efficient to use.

3.1.2 | Comparison of PMM and blending factor = 1

In both blended metrics, a blending factor of 1 indicates that full weight is given to the predictive distance. Therefore, using a blending factor of 1 should yield results identical to those obtained by PMM. Even though the results in Table A1, Table A2 and Table A5 are similar, they are not identical. Is this due to the fact that the matcher function is used in the pmm function, but not in the blended function? In some cases, particularly in the MCAR conditions, the both the ranked and scaled versions of the blended metric with a blending factor of 1 perform slightly better than the predictive metric.

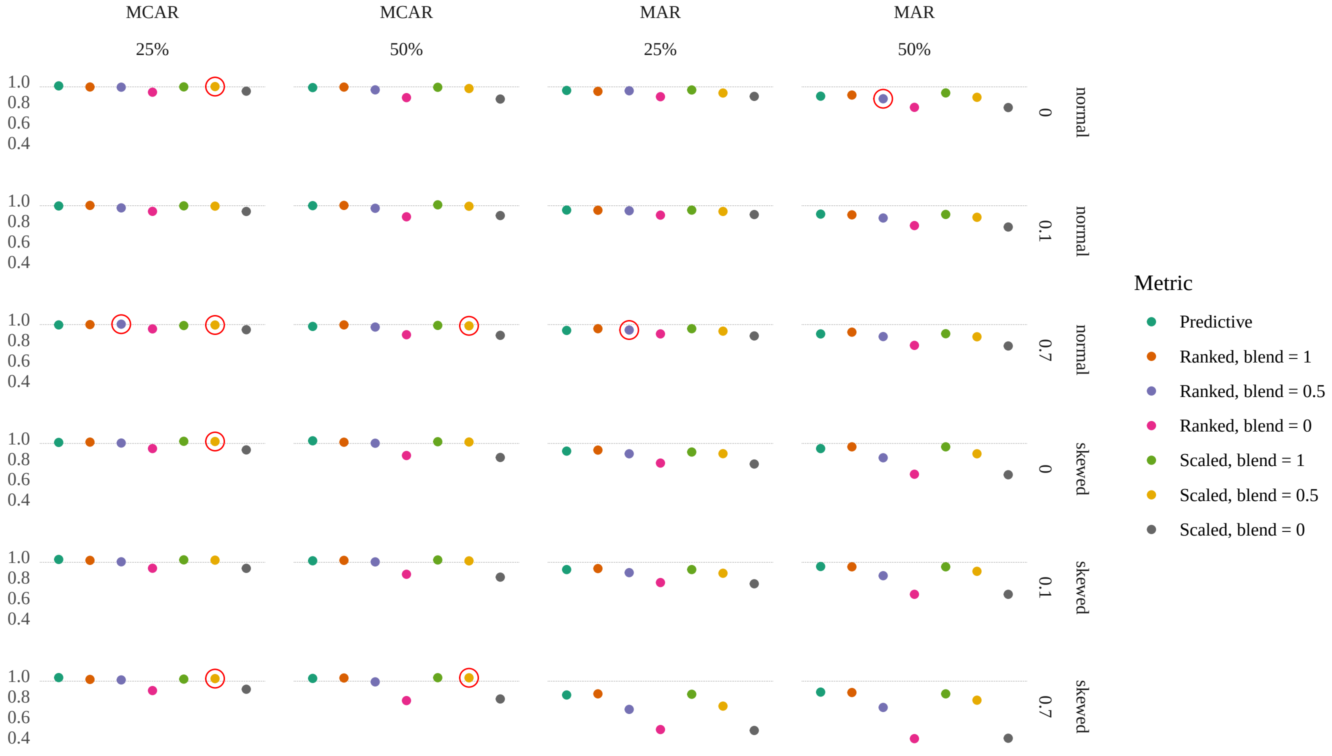

FIGURE 2 Coverage results per condition, where each individual plot shows the results for the seven methods. A reference line is given at coverage = 0.95. Above the plots, the condition combinations of missingness mechanism (MCAR, MAR right) and missingness proportion (25%, 50%) are given. On the right, the condition combinations of distribution (normal, skewed) and correlation (0, 0.1, 0.7) are given. Cases where a smaller blending factor results in higher coverage than either the predictive metric, a blending factor of 1, or both achieve, are circled in red.

3.1.3 | Effect of data generation conditions on performance

In the data generating models, the missingness mechanisms, proportions, skewness of the data, and correlation in the data were varied, resulting in 24 different simulation conditions. The impact of each of these conditions on the performance of the metrics in terms of coverage and bias is mostly as expected, and can be derived from the plots displayed in Figure 2 and 3. The MCAR conditions show higher performance when compared to the MAR right conditions, and a higher proportion of missingness in the data leads to lower performance. The skewness of the data does not always impact the performance negatively. Under the MCAR conditions, a skewed distribution of the data results in higher coverage rates for some cases when compared to a normal distribution. Under the MAR right conditions, however, the opposite is true. Finally, a higher correlation in the data under MCAR conditions does not lead to decreased performance, and in some cases to increased performance. It does lead to lower performance under the MAR right conditions.

3.1.4 | Effect of blending on performance

The effect of the blending factor on performance of the metric can be evaluated for both the ranked version and the scaled version of the metric. For the ranked version, the coverage is almost always higher and the bias almost always smaller when the blended metric is weighted more towards the predictive distance. For the scaled version, the results are similar: in most cases, a higher blending factor leads to better performance.

There are a few exceptions where a blended metric with a blending factor of 0.5 performs better in terms of coverage, which are circled in red in Figure 2. Or is it better to omit the circles? The ranked metric with blending factor = 0.5 performs better than both the predictive metric and the ranked metric with blending factor = 1 in the condition of MCAR with 25% missingness, a normal distribution and a correlation of 0.7. It performs better than the ranked metric with blending factor = 1, but not the

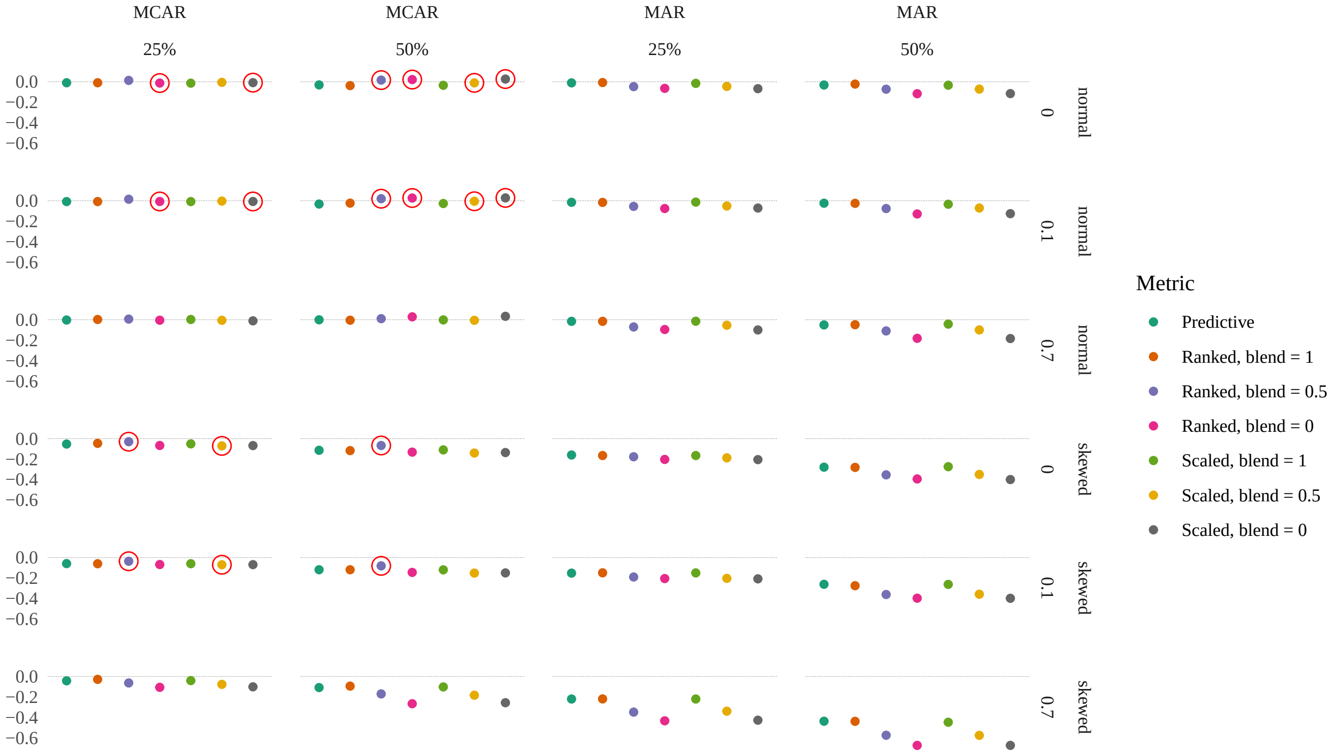

FIGURE 3 Bias results per condition, where each individual plot shows the results for the seven methods. A reference line is given at bias = 0. Above the plots, the condition combinations of missingness mechanism (MCAR, MAR right) and missingness proportion (25%, 50%) are given. On the right, the condition combinations of distribution (normal, skewed) and correlation (0, 0.1, 0.7) are given. Cases where a smaller blending factor results in smaller bias than either the predictive metric, a blending factor of 1, or both achieve, are circled in red.

predictive distance, in the condition of MAR right with 50% missingness, a normal distribution and correlation of 0. Finally, it performs better than the predictive metric but not the ranked metric with blending factor = 1 in the condition of MAR right with 25% missingness, a normal distribution and correlation of 0.7.

The scaled metric with a blending factor of 0.5 outperforms both the predictive metric and the scaled metric with blending factor = 1 in the coverage rate in some cases as well. It performs better than the scaled metric with blending factor = 1, but not the predictive metric, in the conditions of MCAR with 25% missingness and a normal distribution with both a correlation of 0 and a correlation of 0.7, and a skewed distribution with a correlation of 0.7. It performs better than the predictive metric, but not the scaled metric with blending factor = 1, in the conditions of MCAR with 25% missingness, a skewed distribution and correlation of 0, MCAR with 50% missingness, a normal distribution and a correlation of 0.7, and finally, MCAR with 50% missingness, a skewed distribution and a correlation of 0.7.

Additionally, there are exceptions where a blended metric with a blending factor of 0.5 or 0 performs better in terms of bias, circled in red in Figure 3. The ranked metric with blending factor = 0.5 has a smaller bias than both the predictive metric and the ranked metric with blending factor = 1 in the MCAR conditions with 25% missingness, a skewed distribution and a correlation of 0 or 0.1, with 50% missingness, with both the normal and skewed distributions and a correlation of 0 or 0.1.

The scaled metric with a blending factor of 0.5 has a smaller bias than both the predictive metric and the scaled metric with blending factor = 1 in the MCAR conditions with 25% missingness, a skewed distribution and a correlation of 0 or 0.1, with 50% missingness, with a normal distribution and a correlation of 0 or 0.1.

The ranked metric with a blending factor of 0 has a smaller bias than both the predictive metric and the ranked metric with blending factor = 1 in the MCAR conditions with 25% missingness, a normal distribution and a correlation of 0.1, and with 50% missingness, a normal distribution and a correlation of 0. It performs better than the ranked metric with blending factor = 1, but not the predictive metric, in the condition MCAR with 25% missingness, a normal distribution and correlation of 0. Finally, it performs better than the predictive metric, but not the ranked metric with blending factor = 1, in the condition of MCAR with 50% missingness, a normal distribution, and correlation of 0.1.

The scaled metric with a blending factor of 0 has a smaller bias than both the predictive metric and the ranked metric with blending factor = 1 in the MCAR conditions with 25% missingness, a normal distribution and a correlation of 0 or 0.1, and with 50% missingness, a normal distribution, and a correlation of 0. It also performs better than the predictive distance, but not the scaled metric with blending factor = 1, in the condition of MCAR with 50% missingness, a normal distribution, and correlation of 0.1.

3.2 | Simulation study II

3.3 | Application to empirical data

4 | DISCUSSION

This study investigated the properties of a blended metric through simulations and an application to empirical data. In simulation study I, seven metrics were compared: the predictive metric, a ranked version of the blended metric with blending factors of 1, 0.5, and 0, respectively, and a scaled version of the blended metric with blending factors of 1, 0.5, and 0, respectively. The data-generating mechanisms were varied in their missingness mechanism, missingness proportion, distribution, and correlation. A full-factorial design was used, where all possible combinations of metrics and data-generating mechanisms were simulated. The purpose of this study was to investigate whether a ranked and scaled version of the blended metric would yield identical results, a blending factor of 1 would perform the same as PMM, how performance is related to missingness mechanism, proportion, distribution, and correlation, and if blending reduces predictability. The results show that the ranked and scaled versions do not yield identical, but similar results. The scaled version might be favourable to implement because of its computational efficiency. A blending factor of 1 does not yield results identical to those obtained by PMM, but they are similar as well. Performance is higher when the missingness mechanism is MCAR as opposed to MAR right and when the missingness proportion is 25% as opposed to 50%, as would be expected. The skewness of the data does not always impact the performance negatively. Under the MCAR conditions, a skewed distribution of the data results in higher coverage rates for some cases when compared to a normal distribution. Under the MAR right conditions, however, the opposite is true. A higher correlation in the data under MCAR conditions does not lead to decreased performance, and in some cases to increased performance. It does lead to lower performance under the MAR right conditions. Finally, a smaller blending factor, meaning that the metric gives more weight to the Mahalanobis distance, generally leads to lower performance. However, a smaller blending factor results in higher performance in some of the MCAR conditions. It is important to note that these differences are very small, and thus not a reason to suggest that the blended metric should replace the predictive distance under these conditions. It does imply, however, that the blended metric could be implemented in situations where these conditions are true, without compromising in performance. This might be useful to do in cases where it is difficult to choose a particular future time point to match on, or where a user is more interested in similarity between the trajectories of the donors and the target.

just some draft suggestions The current study investigated the influence of missingness proportion, missingness mechanisms, skewness of the data, and correlation in the data for the use of the blended metric. For further investigations of the properties of the blended metric, other factors could be varied, such as the sample size and the number of k matched donors. In addition, it would be interesting to evaluate alternative combinations of similarity measures and the predictive metric. Examples of such measures would be the Frechet distance,[11] and the locally supervised metric learning (LSML) measure.[12] Finally, this study solely used the blending factors of 1, 0.5 and 0, and further research could determine what the optimal blending factor is to predict outcomes.

ACKNOWLEDGMENTS

SUPPORTING INFORMATION

The scripts used for the simulation study are available in the research archive of this study. The study was approved by the Ethical Review Board of the Faculty of Social and Behavioural Sciences of Utrecht University. The approval is based on the documents sent by the researchers as requested in the form of the Ethics committee and filed under number 21-1906.

References

1. Straatmann VS, Pearce A, Hope S, et al. How well can poor child health and development be predicted by data collected in early childhood?. *J Epidemiol Community Health* 2018; 72(12): 1132–1140. Publisher: BMJ Publishing Group Ltd.
2. Cordeiro JR, Postolache O, Ferreira JC. Child's target height prediction evolution. *Applied Sciences* 2019; 9(24): 5447. Publisher: Multidisciplinary Digital Publishing Institute.
3. Van Buuren S. Curve matching: a data-driven technique to improve individual prediction of childhood growth. *Annals of Nutrition and Metabolism* 2014; 65(2-3): 227–233. Publisher: Karger Publishers.
4. Van Buuren S. Broken stick model for irregular longitudinal data. *Journal of Statistical Software* 2020; Submitted for publication: 1–47.
5. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019; 38(11): 2074–2102. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8086doi: 10.1002/sim.8086
6. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 2011; 45(3): 1–67.
7. Herngreen WP, Van Buuren S, Van Wieringen JC, Reerink JD, Verloove-Vanhorick SP, Ruys JH. Growth in length and weight from birth to 2 years of a representative sample of Netherlands children (born in 1988–89) related to socioeconomic status and other background characteristics. *Annals of human biology* 1994; 21(5): 449–463. Publisher: Taylor & Francis.
8. Vink G, Frank LE, Pannekoek J, Van Buuren S. Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica* 2014; 68(1): 61–90. Publisher: Wiley Online Library.
9. Little RJ, Rubin DB. *Statistical analysis with missing data*. 793. John Wiley & Sons . 2019.
10. R Core Team . *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing . 2021.
11. Eiter T, Mannila H. Computing discrete Fréchet distance. tech. rep., Citeseer; 1994.
12. Ng K, Sun J, Hu J, Wang F. Personalized predictive modeling and risk factor identification using patient similarity. *AMIA Summits on Translational Science Proceedings* 2015; 2015: 132. Publisher: American Medical Informatics Association.

How to cite this article: Fopma A.M (2022), A blended distance to define "people-like-me", *Statistics in Medicine*, *x;x:x–x*.

APPENDIX

A RESULTS OF SIMULATION STUDY I

TABLE A1 Method PMM.

mech	mis	dist	cor	qbar	se	t	df	b	2.5%	97.5%	true	cov	bias
MCAR	25%	normal	0	30.46	0.194	0.155	95.417	0.036	29.92	31	30.47	0.96	-0.009
			0.1	30.467	0.192	0.155	94.598	0.035	29.933	31	30.475	0.948	-0.008
			0.7	30.499	0.192	0.163	105.407	0.035	29.967	31.031	30.502	0.947	-0.003
		skewed	0	3.203	0.241	0.183	68.159	0.056	2.534	3.872	3.256	0.96	-0.053
			0.1	3.234	0.254	0.191	59.57	0.062	2.528	3.939	3.292	0.978	-0.059
			0.7	4.067	0.259	0.213	66.255	0.064	3.349	4.785	4.109	0.985	-0.042
	50%	normal	0	30.439	0.337	0.242	32.494	0.109	29.502	31.376	30.47	0.943	-0.03
			0.1	30.443	0.336	0.241	31.331	0.106	29.511	31.374	30.475	0.951	-0.032
			0.7	30.501	0.325	0.243	37.16	0.101	29.599	31.402	30.502	0.932	-0.001
		skewed	0	3.142	0.466	0.378	18.949	0.218	1.848	4.435	3.256	0.976	-0.114
			0.1	3.173	0.478	0.393	20.103	0.231	1.847	4.5	3.292	0.965	-0.119
			0.7	4.001	0.514	0.445	16.984	0.258	2.573	5.429	4.109	0.977	-0.108
MAR	25%	normal	0	30.46	0.19	0.153	97.568	0.034	29.932	30.988	30.47	0.915	-0.01
			0.1	30.461	0.186	0.153	111.053	0.033	29.944	30.977	30.475	0.908	-0.015
			0.7	30.485	0.185	0.161	113.059	0.033	29.972	30.999	30.502	0.893	-0.016
		skewed	0	3.096	0.196	0.155	96.594	0.037	2.55	3.641	3.256	0.875	-0.16
			0.1	3.14	0.195	0.156	96.121	0.036	2.598	3.683	3.292	0.879	-0.152
			0.7	3.89	0.192	0.17	110.599	0.035	3.355	4.424	4.109	0.815	-0.22
	50%	normal	0	30.438	0.299	0.216	41.423	0.087	29.607	31.269	30.47	0.859	-0.031
			0.1	30.452	0.302	0.22	41.664	0.089	29.613	31.29	30.475	0.868	-0.024
			0.7	30.451	0.308	0.231	43.615	0.092	29.596	31.305	30.502	0.859	-0.051
		skewed	0	2.977	0.373	0.273	25.449	0.135	1.941	4.012	3.256	0.9	-0.279
			0.1	3.031	0.373	0.272	25.987	0.134	1.995	4.067	3.292	0.909	-0.261
			0.7	3.672	0.366	0.276	28.879	0.129	2.656	4.689	4.109	0.843	-0.437

TABLE A2 Method ranked, blend = 1.

mech	mis	dist	cor	qbar	se	t	df	b	2.5%	97.5%	true	cov	bias
MCAR	25%	normal	0	30.46	0.194	0.154	96.64	0.036	29.923	30.998	30.47	0.949	-0.009
			0.1	30.467	0.193	0.156	98.919	0.036	29.931	31.003	30.475	0.953	-0.008
			0.7	30.504	0.19	0.162	106.733	0.034	29.976	31.032	30.502	0.95	0.002
		skewed	0	3.21	0.243	0.184	66.597	0.057	2.535	3.884	3.256	0.963	-0.046
			0.1	3.233	0.25	0.188	61.922	0.06	2.54	3.926	3.292	0.969	-0.06
			0.7	4.081	0.255	0.211	74.294	0.063	3.375	4.788	4.109	0.967	-0.028
	50%	normal	0	30.432	0.338	0.242	29.77	0.109	29.494	31.37	30.47	0.948	-0.037
			0.1	30.452	0.342	0.248	30.842	0.112	29.503	31.401	30.475	0.953	-0.023
			0.7	30.496	0.338	0.252	34.916	0.11	29.557	31.436	30.502	0.947	-0.005
		skewed	0	3.139	0.466	0.376	20.963	0.217	1.844	4.433	3.256	0.962	-0.117
			0.1	3.174	0.486	0.397	17.645	0.233	1.825	4.523	3.292	0.969	-0.119
			0.7	4.015	0.513	0.441	17.814	0.254	2.59	5.441	4.109	0.981	-0.094
MAR	25%	normal	0	30.462	0.184	0.151	107.509	0.032	29.95	30.974	30.47	0.906	-0.007
			0.1	30.459	0.19	0.155	100.667	0.034	29.932	30.986	30.475	0.906	-0.016
			0.7	30.485	0.188	0.162	111.273	0.033	29.965	31.006	30.502	0.91	-0.016
		skewed	0	3.091	0.195	0.154	96.209	0.036	2.55	3.632	3.256	0.885	-0.165
			0.1	3.143	0.198	0.157	92.244	0.037	2.594	3.693	3.292	0.889	-0.149
			0.7	3.891	0.194	0.17	108.295	0.036	3.352	4.429	4.109	0.826	-0.219
	50%	normal	0	30.448	0.3	0.216	42.326	0.087	29.616	31.28	30.47	0.87	-0.022
			0.1	30.451	0.302	0.219	41.729	0.088	29.612	31.289	30.475	0.861	-0.025
			0.7	30.452	0.303	0.226	41.153	0.088	29.61	31.295	30.502	0.876	-0.049
		skewed	0	2.975	0.389	0.286	23.008	0.145	1.895	4.055	3.256	0.917	-0.281
			0.1	3.017	0.377	0.278	25.579	0.138	1.971	4.064	3.292	0.906	-0.275
			0.7	3.672	0.362	0.274	30.488	0.127	2.666	4.678	4.109	0.839	-0.438

TABLE A3 Method ranked, blend = 0.5.

mech	mis	dist	cor	qbar	se	t	df	b	2.5%	97.5%	true	cov	bias
MCAR	25%	normal	0	30.484	0.179	0.148	108.198	0.03	29.986	30.981	30.47	0.947	0.014
			0.1	30.49	0.179	0.15	113.573	0.031	29.993	30.986	30.475	0.929	0.015
			0.7	30.508	0.181	0.158	114.675	0.031	30.005	31.011	30.502	0.954	0.006
		skewed	0	3.225	0.202	0.161	92.583	0.039	2.665	3.786	3.256	0.954	-0.03
			0.1	3.257	0.211	0.166	86.542	0.043	2.671	3.843	3.292	0.955	-0.036
			0.7	4.047	0.223	0.191	89.877	0.047	3.428	4.666	4.109	0.962	-0.063
	50%	normal	0	30.486	0.298	0.213	39.727	0.084	29.66	31.313	30.47	0.921	0.017
			0.1	30.495	0.295	0.213	42.344	0.083	29.675	31.315	30.475	0.925	0.02
			0.7	30.512	0.299	0.223	44.219	0.086	29.683	31.341	30.502	0.926	0.01
		skewed	0	3.189	0.357	0.26	29.231	0.122	2.198	4.18	3.256	0.952	-0.067
			0.1	3.211	0.372	0.274	27.569	0.134	2.179	4.243	3.292	0.954	-0.081
			0.7	3.939	0.411	0.326	25.368	0.162	2.797	5.082	4.109	0.943	-0.17
MAR	25%	normal	0	30.422	0.175	0.146	115.774	0.029	29.937	30.907	30.47	0.912	-0.047
			0.1	30.42	0.174	0.148	119.056	0.029	29.935	30.904	30.475	0.901	-0.055
			0.7	30.431	0.174	0.156	128.262	0.029	29.947	30.915	30.502	0.897	-0.071
		skewed	0	3.079	0.176	0.146	113.727	0.029	2.589	3.568	3.256	0.849	-0.177
			0.1	3.103	0.18	0.148	112.696	0.031	2.604	3.601	3.292	0.849	-0.19
			0.7	3.761	0.175	0.157	125.708	0.029	3.275	4.247	4.109	0.674	-0.348
	50%	normal	0	30.398	0.27	0.195	50.044	0.07	29.649	31.147	30.47	0.834	-0.072
			0.1	30.399	0.267	0.195	53.472	0.069	29.658	31.14	30.475	0.83	-0.076
			0.7	30.391	0.271	0.205	53.83	0.07	29.638	31.145	30.502	0.833	-0.11
		skewed	0	2.902	0.29	0.205	43.184	0.08	2.096	3.708	3.256	0.81	-0.354
			0.1	2.931	0.302	0.214	38.277	0.087	2.093	3.769	3.292	0.819	-0.361
			0.7	3.535	0.295	0.217	43.593	0.083	2.715	4.355	4.109	0.693	-0.574

TABLE A4 Method ranked, blend = 0.

mech	mis	dist	cor	qbar	se	t	df	b	2.5%	97.5%	true	cov	bias
MCAR	25%	normal	0	30.458	0.147	0.135	155.345	0.02	30.05	30.866	30.47	0.898	-0.012
			0.1	30.468	0.147	0.137	161.92	0.021	30.06	30.877	30.475	0.895	-0.007
			0.7	30.496	0.147	0.144	168.373	0.02	30.089	30.904	30.502	0.908	-0.005
		skewed	0	3.188	0.148	0.138	158.777	0.021	2.779	3.598	3.256	0.9	-0.067
			0.1	3.224	0.147	0.139	161.171	0.02	2.816	3.632	3.292	0.892	-0.068
			0.7	4.003	0.146	0.157	186.516	0.02	3.598	4.409	4.109	0.858	-0.106
	50%	normal	0	30.492	0.211	0.159	76.07	0.042	29.907	31.077	30.47	0.845	0.022
			0.1	30.502	0.211	0.161	81.039	0.042	29.916	31.088	30.475	0.842	0.027
			0.7	30.529	0.21	0.168	82.535	0.042	29.945	31.113	30.502	0.851	0.028
		skewed	0	3.124	0.208	0.16	82.549	0.041	2.547	3.701	3.256	0.832	-0.132
			0.1	3.148	0.208	0.161	83.064	0.041	2.57	3.725	3.292	0.833	-0.145
			0.7	3.843	0.211	0.18	92.605	0.042	3.257	4.43	4.109	0.76	-0.266
MAR	25%	normal	0	30.406	0.152	0.137	149.215	0.022	29.983	30.829	30.47	0.854	-0.064
			0.1	30.399	0.152	0.138	149.587	0.022	29.976	30.822	30.475	0.858	-0.076
			0.7	30.405	0.15	0.145	160.097	0.021	29.988	30.823	30.502	0.859	-0.096
		skewed	0	3.053	0.144	0.133	160.21	0.019	2.653	3.452	3.256	0.758	-0.203
			0.1	3.088	0.146	0.135	158.768	0.02	2.682	3.494	3.292	0.752	-0.205
			0.7	3.677	0.149	0.146	164.622	0.021	3.264	4.089	4.109	0.477	-0.433
	50%	normal	0	30.354	0.209	0.16	79.339	0.041	29.772	30.935	30.47	0.75	-0.116
			0.1	30.346	0.212	0.163	80.455	0.043	29.758	30.935	30.475	0.756	-0.129
			0.7	30.319	0.213	0.171	82.611	0.043	29.728	30.91	30.502	0.747	-0.182
		skewed	0	2.862	0.206	0.155	83.907	0.041	2.289	3.435	3.256	0.649	-0.394
			0.1	2.895	0.2	0.153	85.405	0.038	2.339	3.451	3.292	0.637	-0.397
			0.7	3.436	0.203	0.163	91.247	0.039	2.872	4.001	4.109	0.388	-0.673

TABLE A5 Method scaled, blend = 1.

mech	mis	dist	cor	qbar	se	t	df	b	2.5%	97.5%	true	cov	bias
MCAR	25%	normal	0	30.457	0.192	0.154	101.05	0.035	29.922	30.991	30.47	0.95	-0.013
			0.1	30.467	0.193	0.155	100.139	0.035	29.933	31.002	30.475	0.949	-0.008
			0.7	30.503	0.188	0.161	110.212	0.033	29.982	31.024	30.502	0.941	0.002
		skewed	0	3.204	0.242	0.184	66.359	0.057	2.531	3.877	3.256	0.971	-0.051
			0.1	3.232	0.253	0.191	61.067	0.062	2.529	3.935	3.292	0.974	-0.06
			0.7	4.069	0.256	0.212	69.564	0.063	3.358	4.78	4.109	0.97	-0.04
	50%	normal	0	30.436	0.337	0.241	32.104	0.108	29.5	31.372	30.47	0.946	-0.034
			0.1	30.448	0.349	0.252	29.01	0.116	29.479	31.417	30.475	0.959	-0.027
			0.7	30.499	0.331	0.245	34.16	0.103	29.581	31.418	30.502	0.942	-0.002
		skewed	0	3.145	0.467	0.379	18.182	0.22	1.848	4.443	3.256	0.967	-0.11
			0.1	3.173	0.474	0.384	18.012	0.223	1.856	4.489	3.292	0.973	-0.12
			0.7	4.007	0.518	0.447	17.319	0.26	2.57	5.445	4.109	0.984	-0.102
MAR	25%	normal	0	30.455	0.188	0.152	101.202	0.034	29.932	30.978	30.47	0.92	-0.015
			0.1	30.462	0.184	0.152	108.357	0.032	29.951	30.973	30.475	0.907	-0.013
			0.7	30.486	0.186	0.161	113.767	0.033	29.97	31.002	30.502	0.91	-0.015
		skewed	0	3.091	0.194	0.154	95.687	0.036	2.552	3.631	3.256	0.866	-0.165
			0.1	3.143	0.196	0.156	93.717	0.036	2.598	3.687	3.292	0.879	-0.15
			0.7	3.889	0.194	0.171	109.694	0.036	3.35	4.429	4.109	0.822	-0.22
	50%	normal	0	30.437	0.306	0.22	38.054	0.09	29.587	31.286	30.47	0.891	-0.033
			0.1	30.441	0.304	0.221	42.355	0.09	29.595	31.286	30.475	0.865	-0.034
			0.7	30.459	0.306	0.229	42.439	0.09	29.608	31.309	30.502	0.861	-0.043
		skewed	0	2.981	0.38	0.277	22.944	0.138	1.926	4.037	3.256	0.917	-0.274
			0.1	3.03	0.368	0.27	28.058	0.132	2.009	4.052	3.292	0.906	-0.262
			0.7	3.662	0.367	0.276	28.672	0.129	2.644	4.681	4.109	0.826	-0.447

TABLE A6 Method scaled, blend = 0.5.

mech	mis	dist	cor	qbar	se	t	df	b	2.5%	97.5%	true	cov	bias
MCAR	25%	normal	0	30.464	0.188	0.152	98.91	0.033	29.943	30.986	30.47	0.953	-0.005
			0.1	30.472	0.184	0.152	111.13	0.032	29.961	30.982	30.475	0.946	-0.003
			0.7	30.496	0.178	0.157	118.072	0.03	30.001	30.99	30.502	0.946	-0.006
		skewed	0	3.185	0.244	0.185	66.797	0.059	2.507	3.863	3.256	0.969	-0.071
			0.1	3.222	0.246	0.186	64.231	0.058	2.54	3.905	3.292	0.971	-0.07
			0.7	4.033	0.259	0.212	68.714	0.064	3.315	4.751	4.109	0.975	-0.077
	50%	normal	0	30.46	0.314	0.224	34.715	0.093	29.587	31.333	30.47	0.934	-0.01
			0.1	30.47	0.303	0.218	39.061	0.088	29.629	31.312	30.475	0.945	-0.005
			0.7	30.496	0.301	0.226	44.523	0.087	29.66	31.333	30.502	0.938	-0.006
		skewed	0	3.114	0.461	0.371	21.338	0.214	1.835	4.393	3.256	0.964	-0.141
			0.1	3.14	0.473	0.387	21.605	0.226	1.826	4.454	3.292	0.964	-0.152
			0.7	3.926	0.506	0.43	18.033	0.247	2.522	5.331	4.109	0.983	-0.183
MAR	25%	normal	0	30.424	0.178	0.149	113.817	0.03	29.929	30.92	30.47	0.89	-0.045
			0.1	30.425	0.179	0.15	113.528	0.03	29.928	30.921	30.475	0.894	-0.051
			0.7	30.448	0.18	0.159	120.713	0.031	29.948	30.948	30.502	0.886	-0.054
		skewed	0	3.068	0.182	0.149	106.933	0.032	2.562	3.574	3.256	0.85	-0.188
			0.1	3.09	0.185	0.151	105.347	0.032	2.577	3.602	3.292	0.843	-0.203
			0.7	3.77	0.185	0.162	114.711	0.032	3.258	4.283	4.109	0.706	-0.339
	50%	normal	0	30.399	0.277	0.2	45.257	0.073	29.629	31.169	30.47	0.848	-0.071
			0.1	30.404	0.282	0.206	48.55	0.077	29.621	31.188	30.475	0.837	-0.071
			0.7	30.402	0.278	0.211	53.383	0.075	29.631	31.173	30.502	0.831	-0.1
		skewed	0	2.906	0.328	0.234	34.28	0.104	1.996	3.816	3.256	0.849	-0.35
			0.1	2.935	0.33	0.235	33.017	0.104	2.017	3.852	3.292	0.862	-0.358
			0.7	3.534	0.347	0.257	32.648	0.115	2.571	4.498	4.109	0.764	-0.575

TABLE A7 Method scaled, blend = 0.

mech	mis	dist	cor	qbar	se	t	df	b	2.5%	97.5%	true	cov	bias
MCAR	25%	normal	0	30.462	0.149	0.135	151.315	0.021	30.049	30.875	30.47	0.908	-0.008
			0.1	30.468	0.146	0.136	156.779	0.02	30.062	30.875	30.475	0.894	-0.007
			0.7	30.491	0.147	0.144	168.358	0.02	30.084	30.898	30.502	0.9	-0.011
		skewed	0	3.188	0.145	0.137	164.684	0.02	2.785	3.591	3.256	0.887	-0.068
			0.1	3.224	0.146	0.138	163.905	0.02	2.818	3.629	3.292	0.891	-0.069
			0.7	4.009	0.149	0.158	177.403	0.021	3.594	4.424	4.109	0.871	-0.101
	50%	normal	0	30.497	0.206	0.158	84.894	0.04	29.925	31.068	30.47	0.831	0.027
			0.1	30.503	0.21	0.16	79.174	0.041	29.92	31.086	30.475	0.854	0.028
			0.7	30.535	0.216	0.171	81.906	0.044	29.934	31.135	30.502	0.845	0.033
		skewed	0	3.12	0.207	0.16	82.676	0.041	2.544	3.696	3.256	0.813	-0.136
			0.1	3.142	0.205	0.16	85.305	0.04	2.572	3.713	3.292	0.805	-0.15
			0.7	3.853	0.211	0.18	94.734	0.042	3.269	4.438	4.109	0.776	-0.256
MAR	25%	normal	0	30.402	0.154	0.138	143.165	0.022	29.975	30.83	30.47	0.857	-0.067
			0.1	30.404	0.154	0.139	145.895	0.022	29.977	30.831	30.475	0.864	-0.071
			0.7	30.401	0.152	0.146	158.825	0.022	29.978	30.825	30.502	0.839	-0.1
		skewed	0	3.051	0.143	0.133	162.883	0.019	2.654	3.448	3.256	0.749	-0.205
			0.1	3.085	0.145	0.135	160.821	0.02	2.682	3.487	3.292	0.74	-0.208
			0.7	3.682	0.149	0.145	163.8	0.021	3.27	4.095	4.109	0.468	-0.427
	50%	normal	0	30.355	0.21	0.16	78.978	0.042	29.772	30.938	30.47	0.748	-0.115
			0.1	30.349	0.209	0.162	84.993	0.041	29.77	30.928	30.475	0.742	-0.126
			0.7	30.317	0.213	0.17	80.768	0.042	29.726	30.909	30.502	0.741	-0.184
		skewed	0	2.856	0.204	0.154	81.374	0.039	2.289	3.422	3.256	0.644	-0.4
			0.1	2.895	0.204	0.155	84.84	0.04	2.328	3.462	3.292	0.637	-0.398
			0.7	3.436	0.207	0.164	84.631	0.04	2.863	4.01	4.109	0.392	-0.673