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

2.1 | Simulation study

This study consists of both simulation research and an application to empirical data. The following paragraphs describe the simulation study in accordance with the ADEMP-structure for reporting simulation studies.[5] The different versions of the blended metric, the aims of the study, the data-generating mechanisms, and the estimand and performance measures are discussed.

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 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.

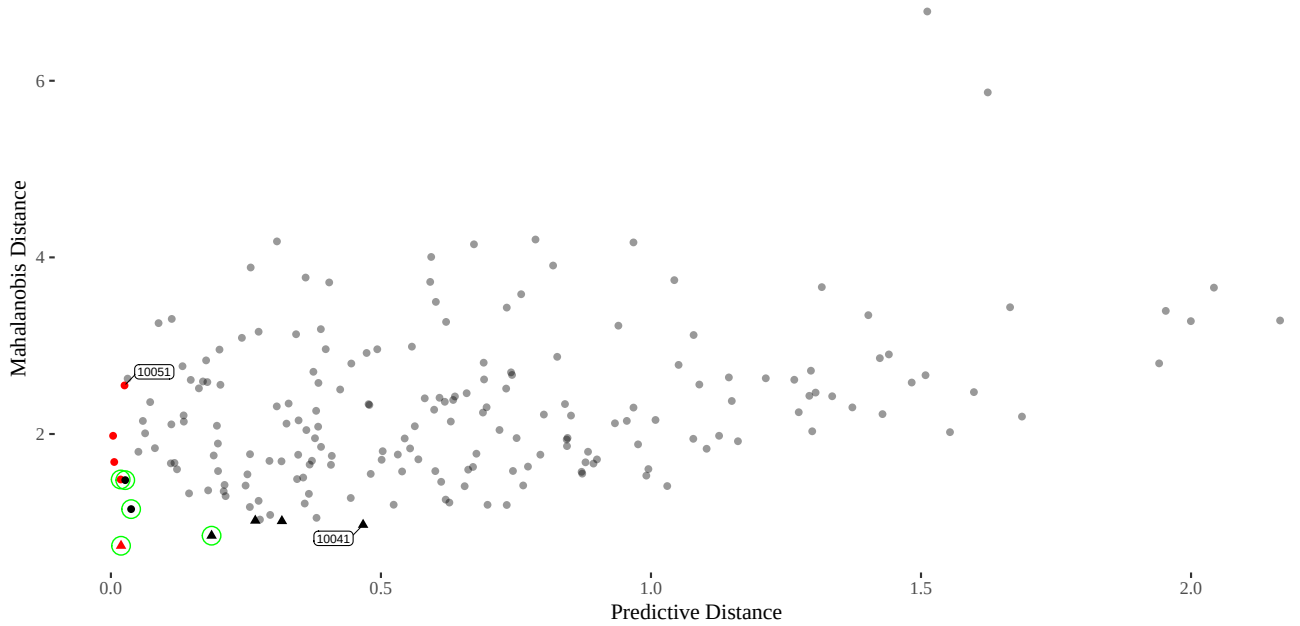

FIGURE 1 Mahalanobis distance plotted against predictive distance for each of the 199 donors. The donors in red are the five matches with the smallest predictive distance, the triangular donors those with the smallest Mahalanobis distance, and the donors circled in green those with the smallest blended 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, weights of respectively 1, 0.75, 0.5, 0.25 and 0 will be evaluated for the blended metric. A weight 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. Using this range of weights allows for an assessment of the degree to which the two distance measures are blended. Using both the ranking and scaling methods and the five different weights, this results in ten different versions of the blended metric to be evaluated.

2.1.2 | Aims

The 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 the ranked and scaled versions of the blended distance measure yield identical results?
2. Does a higher blending factor result in increased similarity between target and matches on the observed predictors, as intended? Should a performance measure like the sum of squared differences be added to check this? Or is the question redundant, as weighting more towards the Mahalanobis distance will automatically lead to smaller differences between the trajectories of the donor and the target?
3. Is there a penalty from blending in terms of reduced predictability?
4. How is the performance of the blended metric related to dimensionality of the data, correlation in the data, and distribution of the data?

It is expected that the ranked and scaled versions do yield identical results, and that a higher blending factor does indeed result in increased similarity between the trajectories of the target and its matches. 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 more weight is given to the Mahalanobis distance. The dimensionality of the data is not expected to have a relevant influence on the performance of the blended metric. 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

In order to answer the previous questions, the blended distance measure will be evaluated in simulated data. All data are generated from one of twelve data-generating mechanisms, with equal means, but with varying dimensionality, variance-covariance matrices, and distributions.

The dimensionality is varied over two conditions. For the first condition, two continuous predictor variables $\mathbf{X_1}$ and $\mathbf{X_2}$ are defined, corresponding to the standardised height measurements (Z-scores) at birth and at 1 month. For the second condition, five continuous predictor variables $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$, $\mathbf{X_4}$ and $\mathbf{X_5}$ are defined, corresponding to the standardised height measurements (Z-scores) at birth, and at 1 month, 2 months, 3 months, and 6 months, respectively. For both conditions, 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 three conditions. The data generating mechanism of the predictor space is a multivariate normal distribution for the first condition, a skewed multivariate normal distribution for the second condition, and a strongly skewed multivariate distribution for the third 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,

$$X = X^2/max\{X\}$$

for the second condition, and

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

for the third condition.

Is this correct? Should only the predictors be skewed, only the outcome, or both?

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

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

and that for the populations with five predictors is given by:

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

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

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 ten different methods to be evaluated and twelve different data-generating mechanisms, the simulation will yield 120 results. From each data-generating mechanism, a sample of size 500 is drawn with a missingness proportion of 0.002, simulating a situation with 499 potential donors and 1 target with a missing height to be predicted. The number of simulations run for each setting is 1000.

2.2 | Estimand and performance measures

The estimand of interest in this study is the predicted height measurement of the target, which in this case is the height at 14 months of age. In both the simulation studies and the application to empirical data, the coverage, confidence interval width, and bias are computed to assess the performance of the blended metric under each combination of conditions. Additionally, the time elapsed for each simulation will be recorded in order to compare the efficiency of the ranked and scaled versions of the blended metric.

2.3 | 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 equivalent to those used in the simulation study: 1, 0.75, 0.5, 0.25, and 0, respectively. 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.

2.4 | Software

R version 4.1.2 (2021-11-01)[9] will be used to simulate the data and perform the analyses. An adaptation of the `mice.impute.pmm` function in the `mice`[6] package 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. The scripts used for the simulation study are available in the research archive of this study.

3 | RESULTS

The simulation results for each setting are given in Table 1, and they are visualised in Figure 2. The figure is just an example I found of what I would like to do for the visualisation of the simulation results. Would this be a good figure/any suggestions

for another way to visualize the results? I'm still in the process of figuring out how to organize the data to get a plot like that in R, but the idea is that performance measures (cov, ciw, bias) are on the right y axis, each combination of dimensionality and correlation on the top x axis (dim = 2 cor = 0.1, dim = 2 cor = 0.7, dim = 5 cor = 0.1, dim = 5 cor = 0.7). Then there would be five boxplots (for each weight, 1, 0.75, 0.5, 0.25, 0) for each distribution (on the bottom x axis, normal, skewed, strongly skewed). The results are discussed below on the basis of the research questions.

3.1 | Ranked vs scaled

Do the ranked and scaled versions of the blended distance measure yield identical results?

3.2 | Blending and historical similarity

Does a higher blending factor result in increased similarity between target and matches on the observed predictors, as intended?

3.3 | Blending and predictability

Is there a penalty from blending in terms of reduced predictability?

3.4 | Influence of dimensionality, correlation, and distribution

How is the performance of the blended metric related to dimensionality of the data, correlation in the data, and distribution of the data?

4 | DISCUSSION

just some draft suggestions The current study investigated the influence of dimensionality, correlation, and skewness of 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 [10], and the locally supervised metric learning (LSML) measure [11] to see what the effects of these factors are. Finally, this study solely used height measurements at certain time points as the predictors in the model, and it would be relevant to include categorical variables in the prediction model as well.

ACKNOWLEDGMENTS

Conflict of interest

The authors declare no potential conflict of interests.

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.

TABLE 1 Simulation results.

Blending	Dimensions	Correlation	Distribution	$p = 1.00$			$p = 0.75$			$p = 0.50$			$p = 0.25$			$p = 0.00$		
				cov	ciw	bias	cov	ciw	bias	cov	ciw	bias	cov	ciw	bias	cov	ciw	bias
Ranked	2	0.1	normal															
			skewed															
			strongly skewed															
		0.7	normal															
			skewed															
			strongly skewed															
	5	0.1	normal															
			skewed															
			strongly skewed															
		0.7	normal															
			skewed															
			strongly skewed															
Scaled	2	0.1	normal															
			skewed															
			strongly skewed															
		0.7	normal															
			skewed															
			strongly skewed															
	5	0.1	normal															
			skewed															
			strongly skewed															
		0.7	normal															
			skewed															
			strongly skewed															

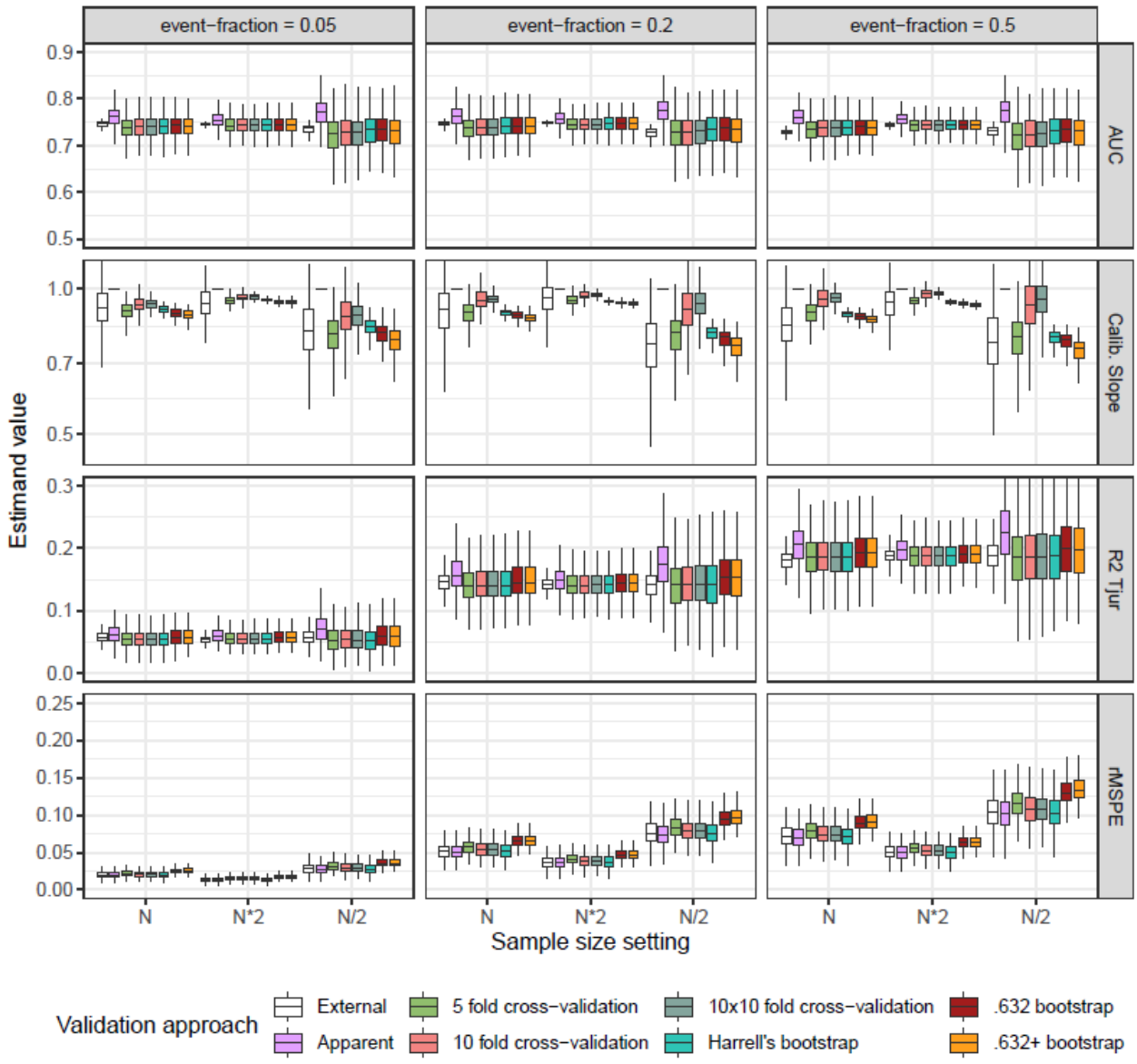

FIGURE 2 Example visualisation of results.

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. R Core Team . *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing . 2021.

10. Eiter T, Mannila H. Computing discrete Fréchet distance. tech. rep., Citeseer; 1994.

11. 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*.