Published in final edited form as:

Med Decis Making. 2018 May ; 38(4): 452–464. doi:10.1177/0272989X17741635.

Matching Microsimulation Risk Factor Correlations to Cross-Sectional Data: The Shortest Distance Method

Sze-chuan Suen, PhD¹, Jeremy D. Goldhaber-Fiebert, PhD², and Sanjay Basu, MD PhD³

¹Epstein Department of Industrial and Systems Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA

²Stanford Health Policy, Centers for Health Policy and Primary Care and Outcomes Research, Stanford University, Stanford, CA, USA

³Stanford Health Policy, Centers for Health Policy and Primary Care and Outcomes Research, Stanford University, Stanford, CA, USA

Abstract

Background—Microsimulation models often compute the distribution of a simulated cohort's risk factors and medical outcomes over time using repeated waves of cross-sectional data. We sought to develop a strategy to simulate how individual's risk factor values remain correlated over time, and compare it to available alternative methods.

Methods—We developed a method using shortest-distance matching for modeling changes in risk factors in individuals over time which preserves both the cohort distribution of each risk factor as well as the cross-sectional correlation between risk factors observed in repeated cross-sectional data. We compared the performance of the method with rank stability and regression methods, using both synthetic data and data from the Framingham Offspring Heart Study (FOHS) to simulate a cohort's atherosclerotic cardiovascular disease (ASCVD) risk.

Results—Correlation between risk factors was better preserved using the shortest distance method than with rank stability or regression (root mean squared difference = 0.077 with shortest distance, versus 0.126 with rank stability and 0.146 with regression in FOHS, and 0.052, 0.426 and 0.352 respectively in the synthetic data). The shortest distance method generated population ASCVD risk estimate distributions indistinguishable from the true distribution in over 99.8% of cases (Kolmogorov-Smirnov $P > 0.05$), outperforming some existing regression methods, which produced ASCVD distributions statistically distinguishable from the true one at the 5% level 15% of the time.

Corresponding Author and Address for Reprints: Sze-chuan Suen, Daniel J. Epstein Department of Industrial and Systems Engineering, 3715 McClintock Ave, Los Angeles, CA 90089, Phone: (213)821-0705, Fax: (213) 740-1120, ssuen@usc.edu.

Application Areas: Global health, HIV and AIDS, Internal Medicine (Cardiology, diabetes, hypertension, infectious disease), Oncology, Public Health

Detailed Methodology: Population Based Studies, mathematical models, decision analysis, discrete event simulation, simulation methods (Monte Carlo Methods, etc.)

Limitations—None of the methods considered could predict individual longitudinal trends without error. The shortest-distance method was not statistically inferior to rank stability or regression methods for predicting individual risk factor values over time in the FOHS.

Conclusions—A shortest distance method may assist in preserving risk factor correlations in microsimulations informed by cross-sectional data.

Keywords

Repeated cross sections; biomarker estimation; microsimulations

Introduction

Microsimulation models track simulated cohorts of individuals in terms of their risk factors and health outcomes over time, and have become increasingly popular for their flexibility and scalability in simulating diseases as varied as cancer, cardiovascular disease, tuberculosis, and HIV [1]–[3]. Microsimulations can not only simulate the cohort average values of risk factors and health outcomes, but also simulate the whole risk factor distribution, which is particularly useful for identifying high-risk cases for intervention targeting. Microsimulations must therefore capture serial correlations in risk factors over time to make accurate predictions about future disease risks or compare the likely effectiveness of alternative interventions. In the case where sufficient longitudinal data is available, statistical methods such as vector autoregressive or multivariate generalized linear mixed models may be used to estimate risk factor distributions over time. Yet in many cases, microsimulation models are not developed from longitudinal data due to the limited availability of such data; hence, short-term longitudinal data or—commonly—repeated cross-sectional data are used [2], [4]–[6]. Microsimulation modelers therefore face the challenge of not observing risk factor changes for individuals directly, and must infer individuals' risk factor values over time and correlations between risk factors at each time period from repeated cross-sectional data or limited longitudinal data.

Inferring correlations between risk factors is particularly important for the accuracy of estimating disease risk in chronic disease models, as multiple risk factors are considered in composite risk scores that can direct treatment choices [7]. For example, current guidelines for treatment of cardiovascular disease using statin medications suggest that physicians use a risk calculator for estimating the 10-year risk of atherosclerotic cardiovascular disease (ASCVD, which is defined as nonfatal or fatal myocardial infarction or stroke) and treating individuals with $\geq 7.5\%$ 10-year risk with a statin; the risk calculation depends on having high combinations of cholesterol, blood pressure, and related risk factors [8]. Error in the correlation between risk factors may mean the over- or under- estimation of these risks, and lead to error in both individual and population forecasts for risk, treatment indication, and treatment cost-effectiveness. Similarly, the most widely-used model for type 2 diabetes, the United Kingdom Prospective Diabetes Study (UKPDS) model, estimates health outcomes (myocardial infarction, congestive heart failure) based on combinations of body mass index, hemoglobin A1c, blood pressure, cholesterol, and other risk factors that imply higher need for treatment among individuals with higher values of risk factors in combination [9]. Diabetes guidelines suggest it is beneficial to direct treatments to the highest-risk

individuals, since treatment itself involves considerable risks from side-effects and the UKPDS model is useful for predicting long-term complications [5].

Microsimulation modelers have commonly used two methods to approximate individual's risk factor correlations over time using repeated cross-sectional data or limited longitudinal data. Some modelers use the 'rank stability' method [2], [4], [7], which assumes individuals' rank in the population for each risk factor remains constant over time. For example, the simulated person with the highest blood pressure in the population will again (subject to survival) have the highest blood pressure one year later, when the blood pressure distribution is sampled from the next age group's distribution. The rank stability method maintains consistency of risk factor values between years, and allows the simulated risk factor distribution to match the observed data at each time period. The rank stability method may not accurately reflect the correlation between risk factors, as the value of each risk factor for each simulated individual at each time point is assigned by rank independently of the others (Figure 1). To our knowledge, the performance of this method has not been systematically evaluated against other methods.

A second method, involving regression [3], [10], relies on the presence of limited longitudinal data. The regression method involves regressing risk factor values at one time period against the values of risk factor values in the previous time period. Since individuals are followed for at least two periods in the data, these regressions capture individuals' trajectories over time. While this autoregressive regression method may perform well at approximating average risk factor values over the simulated population, predicted correlations between risk factors may not reflect those seen in the data.

Here, we compared the rank stability and regression-based methods to an alternative that we developed—the shortest distance method, which assumes an individuals' risk factor values do not vary drastically over time, and uses this assumption to match microsimulations to both cross-sectional correlations and overall risk factor distributions present in limited longitudinal or repeated cross-sectional data over time. We directly compared the performance of the three methods in both simulated and real-world datasets.

Methods

Method Development: The Shortest Distance Method

We developed both unweighted and weighted versions of the *shortest distance method*. The unweighted version uses only cross-sectional data, whereas the weighted version additionally needs two periods of longitudinal data. The k risk factor values for each simulated individual were represented as a k by 1 vector, which can be thought of as a point in space. We sampled individuals from the repeated cross section data at the next time period to create a set of risk factor points that have the same marginal distribution and correlation as seen in the data. We then updated each simulated individuals' risk factor values to these sampled points. Individuals were assigned the point that minimized the overall standardized distance moved across all individuals (Figure 1). This ensures that the population variance and risk factor correlations are preserved for all time periods. The shortest distance method can be thought of as matching each simulated individual to the sampled individual at the

next time period that is “most similar” to him or her by some definition of similarity, and updating the simulated risk factors to the matching person’s risk factors. The matching procedure uses the Munkres algorithm (also known as the Hungarian algorithm), which uses an efficient computational method to find the which points should be matched to minimize the overall distance [11]. Simulation code for replication and application of our algorithm is provided in the Appendix.

Since we drew from the data at each time period, the multivariate distribution of risk factor values will accurately reflect the distribution in the data as long as the sampled population is large enough, and the cross-section correlation between risk factors in the updated simulated population will match that observed in the data (n.b., we use correlation to imply Pearson pairwise correlation between risk factors throughout the remainder of the manuscript). The smallest distance procedure captured time trends in the data, as the sampled risk factors were drawn from subsequent time periods.

We de-mean and standardize the data so each cross section has a mean zero and variance of one for each risk factor, in order to make the definition of distance consistent across multiple risk factors. In the *unweighted* version of the method, we used the Euclidian definition of distance (where the distance between two points x and y is given by $\sqrt{\sum_{j=1}^N (RF_j^x - RF_j^y)^2}$ where RF_j is the j^{th} risk factor value).

However, it may be useful to penalize risk factors that have large variance over time so that they are weighted less heavily than risk factors with smaller variance over time, as risk factors with large variance over time provide less information about how similar individuals are from one time period to the next. If we have access to two waves of longitudinal data, we can use the *weighted* version of the shortest distance method, in which we weight the distance metric by the inverse variance of the risk factor change (distance is expressed as $\sqrt{\sum_{j=1}^N w_j (RF_j^x - RF_j^y)^2}$ where $w_j = 1/\text{var}(RF_{ij}^{t+1} - RF_{ij}^t)$ and RF_{ij}^t is the j^{th} risk factor value for individual i at time t). The equation can be interpreted as a variant of the Mahalanobis distance, where the weights account for the variation across the risk factors over time. We examined both the performance of the unweighted and weighted shortest distance method in our case studies.

Comparison Methods

We compared the unweighted and weighted shortest distance methods to rank stability and regression methods.

For the *rank stability* method, which relies completely on repeated cross section data, risk factor values were sampled from the data at each period. To update values for individuals in the simulation at each period, we first found the rank of each individual’s risk in the cohort (e.g., the person with highest body mass index has rank #1 in the body mass index rank list). The individual with the highest risk factor value in one year at the current period was assigned the highest value sampled for that risk factor in the next period; and the individual with the second highest risk factor value assigned the second highest sample, and so forth.

We also considered the *AR(1) regressions* method for analyses where two periods of longitudinal data are available. For continuous strictly positive risk factors, we performed the ordinary least squares (OLS) regression $\log(RF_k^{t+1}) = C_k + \sum_{j=1}^N \beta_{j,k} \log(RF_j^t) + e_k$ for each risk factor k to estimate constant C_k and coefficients $\beta_{j,k}$, which are then used to predict individual values of RF_k^{t+1} . We then added a simulated Gaussian error term e_k , with zero mean and variance of the residuals to model the variation in the outcome variable to calculate predicted values. These predicted values were treated as the correct values at the next time period, and the same coefficients and constant are used to estimate the subsequent risk factor values. Since we empirically saw that many biomarker distributions have long right tails (as in case study 1, below), we used the logged risk factor values in our regression for better prediction performance [12]. We also explored using square root specifications instead of the log transformation (see Appendix section “Alternative Specifications of the AR(1) Model”). For binary risk factors, values were not logged, and instead a logistic regression was used. Predicted probabilities were then calculated using the logistic distribution function, and these were transformed into binary outcomes where predicted probabilities greater than 0.5 are treated as 1, and 0 otherwise.

For all three methods, we limited our main analysis to individuals known to survive the full period of the simulation, and had complete follow-up over the analysis period, for sake of clarity and to describe comparisons between the methods themselves. We explore the results of a scenario analysis on selective mortality in the Appendix (“Exploring Selective Mortality Effects”).

Comparisons of Method Performance

We used two case studies to compare method performance.

In the first case study, we used biomarker data from the Framingham Offspring Study to compare simulated correlations over time to empirically observed correlations among risk factors. The Framingham Heart Study recruited individuals from Framingham, Massachusetts in 1948 to evaluate risk factors for cardiovascular disease. In 1971, they also began following the offspring of the original cohort, and selected waves of the survey gathered information about all biomarkers listed in Table 1. We used the Offspring Study (instead of the original Framingham cohort) as it provided a larger number of individuals with complete data for all nine biomarkers of interest. For illustrative purposes, we focused on a single cohort of white males between 55–65 years of age at their first observation who had four or more consecutive observations (total analysis periods $T = 4$) and complete entries for all nine risk factors of interest, so all individuals had at least 11 years of fully observed follow-up data ($N = 206$ individuals). While there is variation in age, we treated this group as a single cohort and do not adjust for age in any of the methods. Note that this population is not representative of the general population – by the last time period, the cohort is 66–77 years old – and this population was selected because they have complete risk factor information over four consecutive time periods, and therefore serve only to illustrate the performance of the various methods.

Since the Framingham Offspring is a panel dataset, we could compare the simulated output of each method to the true data to evaluate method performance for individuals over time. To emulate repeated cross-sectional data for use in our analyses, we sampled N individuals from the data to generate “cross sectional” data for each time period. Each “cross section” therefore included data from potentially different individuals, just as a real cross-sectional dataset would, but would reflect the overall risk factor distributions in the true cohort. The sampling processes produced stochastic noise in our results, so we repeated our analysis 10,000 times (each time drawing N individuals) and reported average and variance estimates over these iterations. We also varied the number of individuals sampled to evaluate the performance of the methods with fewer individuals in the cohort (see Appendix section “Effect of Smaller Sample Sizes”).

In the second case study, we used synthetic data to explore performance of the methods when using data with different characteristics. We generated synthetic data by simulating four “risk factors” of 500 individuals over 10 time periods ($T = 10$). We examined model performance when the correlation between risk factors changed over time (which may happen as individuals age [13]) and there are large fluctuations in risk factors values over time (as may be the case if there is feedback between biomarker levels). The code used to generate each of these datasets is included in the Appendix.

Outcome Metrics

For each method, we evaluated two main outcome metrics: (i) the ability of each method to replicate the correlations between risk factors, and (ii) the longitudinal accuracy in risk factor values over time.

We estimated the root mean squared (RMS) correlation error as the square root of the average squared difference between all pairwise correlations of risk factors (not including correlation with a risk factor to itself) between the true and predicted risk factors. For k risk factors, we have a total of $\binom{k}{2}$ pairwise correlations, and we calculate RMS correlation error as:

$$\sqrt{\frac{\sum_{i,j, i \neq j} \text{corr}(RF_i, RF_j)^2}{\binom{k}{2}}}$$

Modelers would also like to accurately capture the risk factor values over time to identify how risk factors change for each individual. We measured the longitudinal accuracy using the RMS error per individual per time period for each risk factor. For individual i at time t , the estimated risk factor RF value ($\widehat{RF}_{i,t}$) and true value ($RF_{i,t}$), T total time periods, and N individuals, the RMS error per individual per time period is given by:

$$\sqrt{\frac{\sum_{t=1}^T \sum_{i=1}^N (\widehat{RF}_{i,t} - RF_{i,t})^2}{N \times T}}$$

We also compare the similarity between the predicted risk factor values at over time to explore whether one prediction method is more likely to over- or under-estimate changes in risk factor in values over time. We calculate the correlation between risk factor values at time period 1 and the last simulated time period for each individual (this measures correlation over time, or autocorrelation, as opposed to the correlation between risk factors for the RMS correlation error, above). We report the population mean and standard deviation of this autocorrelation value for each risk factor.

In the Framingham Offspring case study, we additionally assessed the clinical significance of error and the distribution of health risks across the analysis cohort. We compared the distribution of estimated 10-year atherosclerotic cardiovascular disease (ASCVD) risk scores calculated using the risk factors estimated using each method at the final time period ($T=4$). 10-year ASCVD risk scores were calculated using The American College of Cardiology/American Heart Association (ACC/AHA) Pooled Cohort risk equations to guide statin initiation for non-Hispanic adults aged 40–79 [14]. These equations use the results of Cox proportional hazards equations to estimate 10-year risk of ASCVD from a combination of individual risk factors (age, sex, race, total and high density lipoprotein cholesterol, systolic blood pressure, whether a person is taking blood pressure treatment, smoking status, and type 2 diabetes status [fasting glucose ≥ 126 mg/dL]), based on multiple cohort studies including the Framingham Offspring data we use. Accurate prediction of ASCVD score distributions require accurately modeling the correlation between risk factors. If the distribution estimated using a particular method differed significantly from the true distribution, it would show that the correlation error for that method may be significant enough to bias clinical outcomes.

Since the ASCVD risk scores are often used to distinguish high- and low-risk patients, we evaluated the ability of each method to distinguish patients across cutoff thresholds between high- and low-risk. We found the false positive rate, true positive rate, and accuracy for each cutoff value for each method. We explore more cutoff values than are typically used by clinicians for ASCVD scores (7% and 10%) to gain more general insights about method performance.

We also compared the ASCVD distributions generated from each estimation method using the two-sided, two-sample Kolmogorov-Smirnov (KS) test. This statistical hypothesis test uses the maximum absolute difference between CDFs created from two samples to evaluate the null hypothesis that the samples are drawn from the same distribution. We use the KS test to evaluate whether the CDFs of the ASCVD distribution generated using each method can be distinguished from the distribution generated by the true risk factor values at the last time period.

We also examine scenarios using the Framingham Offspring case study to examine the effect these methods could have on cost-effectiveness and selective mortality (see Appendix).

Since we drew individuals for our analysis from the Framingham Offspring dataset, some stochasticity was introduced into our analysis. We therefore ran 10,000 iterations of our

analysis, estimating each method on all iterations, to ensure our results were not driven by stochastic noise. Simulations were performed in Matlab R2013b.

Results

Framingham Offspring Case Study

Correlation and ASCVD Scores—For identical starting populations, the shortest distance and rank stability methods matched the true risk factor correlations more closely than AR(1) after 11 years. The root mean squared difference in correlations from the true Framingham Offspring data over all risk factors was smallest using the shortest distance method (0.077), with almost twice as much average error using rank stability (0.126) or AR(1) regression (0.146). Figure 2a shows example pairwise correlations for each of the estimation methods (all pairwise correlations are shown in the Appendix). The higher correlation error for the AR(1) regression method is apparent from the pairwise risk factor correlations, as the AR(1) correlation estimates often differ significantly from the true correlations.

Correlation errors may lead to differences in prediction that are significant for ASCVD 10-year risk score estimates. Figure 2b shows that AR(1) performs with lower accuracy for ASCVD cutoff values over 0.25. AR(1) regression's poorer performance on capturing ASCVD risk score estimations were confirmed by the Kolmogorov-Smirnov statistic results (see Table 2), which showed that AR(1) ASCVD risk score distributions were different from the true distribution at the 5% level 15% of the time (out of the 10,000 iterations). By contrast, rank stability was significantly different only 2.6% of the time, and both shortest distance methods less than 0.2% of the time. A hypothetical microsimulation that used the AR(1) method on this data to approximate cohort risk would be less likely to capture the true risk of ASCVD observed in the data. We demonstrate how inaccuracies in predicting composite measures like ASCVD risk scores might influence cost-effectiveness results in the Appendix section "Exploring the Effect on Cost-Effectiveness Analyses."

Longitudinal Error—The average per capita RMS error was similar across all methods, with no method consistently performing the best across all risk factors. Table 3 presents the RMS error per capita error per time period for each risk factor. All three methods performed well for some risk factors. BMI, for example, had average per capita error per time period was less than one standard deviation for all methods, or ranged between 2 and 4 kg/m² across all methods. This is a relatively small value given that the range between 5% and 95% of the true BMI values lie between 22 and 35 kg/m² in the data. However, simulations for other risk factors -- the binary outcomes in particular -- did not perform as well, with error around 0.4 for smoking status (compared to 0.5 error associated with predicting binary outcomes using a coin flip).

Autocorrelation—We found that the shortest distance and rank stability methods do not consistently have higher or lower average autocorrelation values than the true or the AR(1) regression values, and nor does the autocorrelation measure have consistently larger or smaller standard deviations across the population. This demonstrates that while these methods would reasonably be expected to generate patterns with higher autocorrelations due

to their estimation method of preserving either rank stability or distance across time periods, this is not always the case, and the autocorrelation across methods varies with what data is used. Full autocorrelation results are provided in the tables in the “Autocorrelation” section of the Appendix.

Effect of Smaller Sample Sizes—We found that the performance of shortest distance and rank stability methods continued to perform as well or better than AR(1) regression methods with smaller cohort sizes, despite expectations (see Appendix section “Effect of Smaller Sample Sizes” for details). AR(1) continued to perform worse in capturing correlation even if cohort sizes were small. RMS correlation error remained generally constant with an average of 0.08 and 0.13, respectively, for shortest distance and rank stability methods when the analysis cohort sampled was above 50% of the original (N=103). Below 50%, RMS correlation error for these methods rose to averages of 16% and 20%, respectively. However, the RMS correlation error for the simulated values estimated using AR(1) regression remained higher than those for shortest distance or rank stability for all sampled percentages (average correlation error of 0.15 for percentage sampled above 50%, and 21% below 50%). While performance on correlation error suffered if the number of individuals in the cohort were small, shortest distance was still able to outperform other methods. Per capita longitudinal error also did not rise significantly if cohort sizes were small, with error values generally holding steady at values observed when the full cohort was used. Again, we observed no consistent patterns in autocorrelation over methods between the first and last time period even when cohort sizes were small.

Synthetic Data Case Study

Using synthetic data, we found that when correlations between risk factors changed over time, rank stability performed poorly in capturing the changes as compared to regression and shortest distance methods (Figure 3). We increased the correlation between risk factors 1 and 2 over ten time periods (panel a) while risk factors 3 and 4 were sampled randomly from the normal distribution. The rank stability method erroneously estimated the correlations between risk factors 1 and 2 to remain relatively constant over time (panel b), remaining similar to the correlation values observed in the first time period. By contrast, shortest distance methods always maintained high correlation accuracy even when correlations changed (since it did not sample the risk factors independently of each other). This was reflected in the RMS correlation error, which was 0.052 for both shortest distance methods, 0.426 for rank stability, and 0.352 for AR(1) regression.

While it is unsurprising that AR(1) regression performs poorly in capturing correlation, this example shows that the rank stability method is not able to capture changing relationships between risk factors. For example, if an individual happened to have the highest ranked risk factor 1 and the lowest value for risk factor 2 in the first period, the rank stability method assumed that the relationship among the risk factors persisted throughout all of the analysis period. This became a problematic assumption if, over time, those with high values of risk factor 1 also come to have high values of risk factor 2 (correlation between risk factor 1 and 2 increases). Rank stability kept the Spearman’s rank correlation coefficient constant over all

time periods, limiting the method's ability to reflect changes in both Spearman's and Pearson correlation coefficient.

We also tested the effect of large individual fluctuations in risk factor values in an example where two risk factors demonstrate feedback behavior (i.e., as might be exhibited in the case of blood glucose and insulin, or oscillations due to homeostatic patterns). Such fluctuations might be more likely to be observed in data at short time scales (i.e., daily or weekly data). Risk factor 4 follows risk factor 3 with a one period lag, and risk factor 3 oscillates over time (Figure 4a). In this case, the average per person RMS error for both risk factors 3 and 4 are much higher for all methods besides AR(1), which was close to zero error for both risk factors (Figure 4b, circles). The regression AR(1) would be expected to perform best in this case, since it is best able to capture time-dependent capture patterns in the data whereas the rank stability and shortest distance methods rely on consistency over time in risk factor values.

AR(1)'s success was dependent on its ability to capture the entirety of the fluctuating pattern in the data it used for estimating the coefficients (the first two periods of longitudinal data). It performed well if each period's risk factor 3 and 4 predict next period's risk factors 4 and 3. However, if the fluctuation only appeared in one variable, AR(1) did worse than the other methods, since it did not have sufficient information to predict the fluctuations with accuracy. Running this analysis again (Figure 4b, dashed lines), where only risk factor 4 oscillates (risk factor 3 is replaced with Gaussian noise), the average per capita RMS error for AR(1) increased to 18.6. In this case, while the other methods perform similarly to the double risk factor fluctuation case, the AR(1) RMS error oscillated, with a range of 7.7 and 30.8 (contrast to weighted shortest distance, which has a mean per capita RMS error of 9.3 and range (5.0, 14.5)).

Discussion

Microsimulation models often require use of limited longitudinal data or, more commonly, repeated cross-sectional data to simulate risk factors among individuals over time. Particularly in microsimulations of chronic disease risk, multiple risk factors not only need to be accurately estimated in isolation, but their correlations need to be accurately simulated to ensure both individual- and population-level accuracy. As clinical decision-making is increasingly leaning towards use of microsimulation models to inform guidelines and cost-effectiveness analyses [2], [5]–[7], [15], we developed a shortest-distance method for preserving risk factor correlations over time among microsimulations, and compared the method to two commonly-used alternatives.

We found that shortest distance matching methods were able to accurately capture the correlations between multiple risk factors when such risk factor values were not expected to fluctuate widely but still expected to change over time. The shortest distance method was particularly effective compared to rank stability methods when certain risk factor values were expected to change more than others over time (e.g., for cardiovascular risk, blood pressure and cholesterol values tend to change more than smoking status or diabetes status).

Regression methods performed best with widely fluctuating, time-dependent risk factors but not with more typical variations.

In the Framingham Offspring data, regression with AR(1) had the greatest difficulty capturing cross-sectional risk factor correlations at multiple time points, and in general, shortest distance methods outperformed rank stability methods for this metric when correlations changed over time. Shortest distance methods may therefore be preferred if the analysis period is long or includes periods of transition where the correlation between risk factors may change (as may be the case as individuals age). Furthermore, while changes in rank for any single risk factor compromised the performance of rank stability, shortest distance was more robust to fluctuation if there were many other risk factors in the analysis and the remaining risk factors were relatively constant over time (since distances in depend on all the risk factors simultaneously). The results of the cost-effectiveness analysis and mortality example (see Appendix) showed that shortest distance methods can be reasonable choices even in the presence of selective mortality and when trying to estimate population outcomes for cost-effectiveness analyses.

Shortest distance also maintained good performance when an entire cohort's risk factor values increased or decreased over time -- the whole risk factor distribution was translated over time, instead of individuals' risk factors varying widely -- as illustrated in the Framingham Offspring case study, where average diastolic blood pressure steadily decreased over time. Despite its reliance on matching individuals over time, shortest distance methods were also robust to reasonable levels of small sample sizes. The RMS error for each risk factor was similar to those of the AR(1) regression and rank stability even when only 10% of the study population was sampled, and RMS correlation error remained constant until less than 50% of the population was sampled.

Code for implementing our shortest distance methods can be found in many software packages, as any algorithm for solving the assignment problem may be used. For example, assignment problem solvers are available in R [16], Munkres algorithms in Python [17], and versions in Matlab [18]. We share our MATLAB code in the Appendix to this paper.

In some cases, it may also be possible to choose different methods to predict different risk factors, thus using multiple methods in one analysis, if there is external data or data is held out from the prediction to validate each method's outputs. One method may be more accurate at predicting some subset of the risk factors (for instance, if one risk factor has a well-established linear trend with time, AR(1) regression may be most suitable, and if another is such that individuals always remain in the same rank relative to each other, rank stability is most suitable, etc.). Then the most accurate method should be used on each subset of risk factors. The feasibility of this approach will depend on the available data and prior knowledge about how each risk factor changes over time.

Our methods and results have important limitations. First, we found that shortest distance methods may not be the best choice when risk factors are expected to have strong relationships with each other or fluctuate over time, as these patterns may be better captured with prediction methods that explicitly model these trends. If there is sufficient coupling

between variables, AR(1) may be able to outperform shortest distance or rank stability methods by exploiting patterns where one variable predicts the changes in another. However, good predictors of fluctuation may not always be available, as demonstrated in the case where only one risk factor oscillated over time, in which case AR(1) also did not perform well.

Secondly, the shortest distance methods required higher computational resources than either of the other methods. The Munkres assignment algorithm (used in this case for the shortest distance matching method) was of polynomial runtime, with worst-case $O(n^3)$, meaning that as the number of individuals to be matched increased by x , the worst-case computational run time increased by a factor of x^3 . By contrast, the rank stability algorithm needed only to sort individuals by risk factor, and therefore had $O(n \log n)$ run time, and AR(1) regression needed to be estimated once and then evaluated using matrix multiplication at each time step (runtime of $O(nm)$, where m is the number of risk factors). Therefore both rank stability and AR(1) methods had fewer computational requirements than shortest distance methods. With the number of individuals used in the case studies here (up to 500 individuals in the synthetic cohort), all methods were very tractable, with shortest distance methods requiring between 3–5 seconds of run time for each time period (on an Intel i7-6700 CPU 2.6Ghz, 8GB RAM system). However, with a larger simulated population, it may not be computationally feasible to use the shortest distance methods directly.

Even so, a large population may potentially be separated into smaller groups (i.e., those with cholesterol, BMI, etc. between certain ranges) and the shortest distance method could be performed on these smaller subgroups with greater ease. The resulting correlation between risk factors should still reflect the observed population, although error over time for individuals may increase.

Third, as with any simulation model, assumptions about future trends must also be made when forecasting health outcomes. In the case of AR(1) methods, the modeler may assume that the regression coefficients stay constant in the future (i.e., risk factor relationships over time do not change), or otherwise make assumptions about time trends in the mean and variance of risk factors. In the case of shortest distance or rank stability methods, the modeler needs to make assumptions about the risk factor distribution at subsequent time periods. The distribution mean, covariance matrix, and other statistical properties could be estimated using the existing data and projected into the future using a variety of techniques (multivariate linear mixed models, time series analysis approaches, moving averages, Kalman filtering, etc.) or else assumed to remain the same as the last observed time period. Shortest distance and rank stability methods could then be used after sampling from these proposed distributions.

In future work, microsimulation methods can be developed to explore variations in weighting schemes. Here, we have explored one weighting scheme for shortest distance methods, where the weights for each risk factor reflected the variance observed across two periods of longitudinal data. Alternative weighting schemes could be tailored for specific applications. For instance, knowledge about subpopulations could be incorporated. For instance, a modeler may wish to assume that people with diabetes (high HbA1c) have

elevated cholesterol levels that shift over a much greater range than non-diabetics. Then a modeler could have different distance weights for those with low HbA1c and those with high HbA1c measures to allow for these trends. In general, weighting the distance metric allows for greater flexibility to the particular application and set of risk factors at hand.

In addition, other metrics besides Euclidian or Mahalanobis distance or other distance weighting schemes may be considered, which may be desirable if there are nominal or ordinal variables. A variety of distance metrics have been explored in recent years as distance measurements play a key role in many machine learning methods (e.g. clustering). For instance, in the case where there are several nominal or ordinal variables, a modeler may wish to use a matching coefficient or Jaccard distance (a ratio using the number of matches and non-matches between the elements of the vectors). If there is only one nominal or ordinal variable, one could simply look at whether the values matched (distance of 0 or 1) multiplied by a weight. Other measures include the Kendall distance, Minkowski Distance, Chebyshev Distance, and Spearman Distance. Ultimately, there are many ways to implement shortest distance methods, and we encourage modelers to consider them when projecting risk factor values using limited cross-sectional data.

In the current work, we developed a method that captures correlations between risk factors over time even when longitudinal data is not available. The shortest distance method could be valuable for studies where multiple risk factors inform the risk of disease. The shortest distance method outperforms existing methods in capturing correlation even when the correlation changes in successive waves of the cross-sectional data. Because shortest distance may not perform well in cases where individuals' risk factors exhibit large fluctuations, future work remains in developing techniques to appropriately model multifactorial heterogeneity over time when individual data over time is not observed. Shortest distance methods may nevertheless be valuable as individual-level simulations become more popular.

Acknowledgments

Funding Sources: Financial support for this study was provided in part by the NIH's National Institute on Aging (K01 AG037593; PI: Goldhaber-Fiebert), by Stanford's Freeman Spogli Institute's Underdevelopment Action Fund (PI: Goldhaber-Fiebert), by the National Institute On Minority Health And Health Disparities of the National Institutes of Health under Award Numbers DP2MD010478 and U54MD010724 (PI: Basu), and by the National Heart, Lung, And Blood Institute of the National Institutes of Health under Award Number K08HL121056 (PI: Basu). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Birnbaum J, Gadi VK, Markowitz E, Etzioni R. The Effect of Treatment Advances on the Mortality Results of Breast Cancer Screening Trials : A Microsimulation Model. *Ann. Intern. Med.* 2016; 164:236–243. [PubMed: 26756332]
2. Basu S, Yudkin JS, Sussman JB, Millett C, Hayward RA. Alternative Strategies to Achieve Cardiovascular Mortality Goals in China and India : A Microsimulation of Target-Versus Risk-Based Blood Pressure Treatment. *Circulation.* 2016; 133(9):840–848. [PubMed: 26762520]
3. Hontelez JAC, Chang AY, Ogbuaji O, Vlas SJ, Bärnighausen T, Atun R. Changing HIV treatment eligibility under health system constraints in sub-Saharan Africa: Investment needs, population

- health gains, and cost-effectiveness. *AIDS*. 2016; 30:2341–2350. no. October 2015. [PubMed: 27367487]
4. Lim SS, et al. Prevention of cardiovascular disease in high-risk individuals in low-income and middle-income countries: health effects and costs. *Lancet*. 2007; 370(9604):2054–2062. [PubMed: 18063025]
 5. Basu S, Shankar V, Yudkin JS. Comparative effectiveness and cost-effectiveness of treat-to-target versus benefit-based tailored treatment of type 2 diabetes in low-income and middle-income countries. *Lancet Diabetes Endocrinol*. 2016; 4(11):922–932. [PubMed: 27717768]
 6. Sussman J, Vijan S, Hayward R. Using Benefit-Based Tailored Treatment to Improve the Use of Antihypertensive Medications. *Circulation*. 2013; 128(21):2309–2317. [PubMed: 24190955]
 7. Stone NJ, et al. 2013 ACC / AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults. *Circulation*. 2014; 129
 8. Muntner P, et al. Validation of the Atherosclerotic Cardiovascular Disease Pooled Cohort Risk Equations. *Jama*. 2014; 311(14):1406. [PubMed: 24682252]
 9. Hayes AJ, Leal J, Gray AM, Holman RR, Clarke PM. UKPDS Outcomes Model 2: a new version of a model to simulate lifetime health outcomes of patients with type 2 diabetes mellitus using data from the 30 year United Kingdom Prospective Diabetes Study: UKPDS 82. *Diabetologia*. 2013; 56(9):1925–1933. [PubMed: 23793713]
 10. BJ Chen BK, Jalal H, Hashimoto H, Suen SC, Eggleston K, Hurley M, Schoemaker L. Forecasting trends in disability in a super-aging society: Adapting the Future Elderly Model to Japan. *J. Econ. Ageing*. 2016; 8:42–51. [PubMed: 28580275]
 11. Munkres J. Algorithms for Assignment and Transportation Problems. *J. Soc. Ind. Appl. Math*. 1957; 5(1)
 12. Mangiafico, SS. Summary and Analysis of Extension Program Evaluation in R, 1.2.1. New Brunswick, NJ: Rutgers Cooperative Extension; 2016.
 13. Hardy ST, et al. Heterogeneity in Blood Pressure Transitions Over the Life Course Age-Specific Emergence of Racial/Ethnic and Sex Disparities in the United States. *JAMA Cardiol*. 2017:1–9.
 14. Goff DC, et al. 2013 ACC / AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation*. 2014; 129(25)
 15. Vijan S, Sussman JB, Yudkin JS, Hayward RA. Effect of Patients' Risks and Preferences on Health Gains With Plasma Glucose Level Lowering in Type 2 Diabetes Mellitus. *JAMA Internat Med*. 2014; 174(8):1227–1234.
 16. [Accessed: 11-Jan-2017] Solve Linear Sum Assignment Problem. 2017. [Online]. Available: https://artax.karlin.mff.cuni.cz/r-help/library/clue/html/solve_LSAP.html
 17. Karlin Lab. [Accessed: 11-Jan-2017] munkres 1.0.9. Python Software Foundation. 2017. [Online]. Available: <https://pypi.python.org/pypi/munkres/>
 18. Cao, Y. [Accessed: 11-Jan-2017] Hungarian Algorithm for Linear Assignment Problem (V2.3). Matlab Mathworks. 2011. [Online]. Available: <http://www.mathworks.com/matlabcentral/fileexchange/20652-hungarian-algorithm-for-linear-assignment-problems--v2-3->

Appendix

Alternative Specifications of the AR(1) Model

We explored using a square root transform instead of a log transform for the AR(1) model (used in the main analysis). We find that the square root method does not perform as well as the log transform specification, overestimating the 75th percentile and mean of the ASCVD risk score distribution. However, the maximum predicted ASCVD risk score values are similar to the log specification (see box plots of the ASCVD distribution, below). As the log transformation is closer to the true distribution, we use the log transformation for our main analysis.

Box Plot of ASCVD 10-Year Risk Score Distribution

All Pairwise Correlations Between Variables in Framingham Offspring Case Study, by Estimation Method

Autocorrelation

We present the autocorrelation of each biomarker for each estimation method in the table below. The frequency of observation influences autocorrelation – for a biomarker that changes slowly over time, we would expect high autocorrelation if the frequency of observation is high, and lower autocorrelation if the observations are spaced further apart. In

the Framingham Offspring data, the observations were taken 3 to 4 years apart, resulting in relatively low autocorrelation in the true data.

We find that the shortest distance and rank stability methods do not always have lower average autocorrelation values than the true or the AR(1) regression values, and nor does the autocorrelation measure have smaller standard deviations across the population. This demonstrates that while these methods would reasonably be expected to generate patterns with higher autocorrelations, this is not always the case, and the autocorrelation across methods may vary with what data is used.

Autocorrelation Table

Correlation Between Risk Factor Values at Time Period 1 and 4 in the Cohort, by Prediction Method

	True	Shortest Dist.	Shortest Dist.Wtd.	Rank Stability	AR(1) Reg.
BMI					
Mean	0.050	0.051	-0.049	-0.005	-0.029
SD	0.074	0.047	0.043	0.063	0.084
HDL					
Mean	-0.015	0.000	0.007	0.009	0.029
SD	0.053	0.056	0.058	0.039	0.022
TotChol					
Mean	0.031	0.017	-0.028	0.019	-0.003
SD	0.051	0.060	0.043	0.033	0.039
Glu					
Mean	-0.015	0.011	-0.015	-0.048	-0.056
SD	0.032	0.025	0.062	0.030	0.055
Tri					
Mean	0.002	0.021	-0.019	0.020	-0.033
SD	0.044	0.025	0.067	0.053	0.058
BPtrt					
Mean	0.078	-0.027	0.000	-0.028	0.039
SD	0.044	0.105	0.077	0.057	0.034
Smoke					
Mean	0.012	0.025	-0.006	0.025	0.013
SD	0.045	0.061	0.047	0.083	0.049
SBP					
Mean	0.005	-0.031	-0.002	0.004	0.038
SD	0.068	0.080	0.061	0.050	0.049
DBP					
Mean	-0.011	-0.006	0.020	0.004	-0.031
SD	0.095	0.081	0.074	0.061	0.058

Accuracy, True and False Positive Rate in the Framingham Offspring Case Study

We have separately plotted the true positive rate, false positive, and accuracy (the number of true positive plus true negatives as a fraction of all individuals) for each ASCVD cutoff value between 0 and 0.8 (by ASCVD cutoff value we mean the value that separates “high” and “low” ASCVD scores). The shortest distance and rank stability rates perform similarly to each other regardless of cutoff value. AR(1) regression is less accurate than the other methods at higher cutoff values, by at most 5%. This is consistent with the Kolmogorov-Smirnov Statistic Results, which showed that AR(1) regression produced ASCVD risk score distributions that were statistically distinguishable from the true distribution more than the other methods.

The Effect of Smaller Sample Sizes

Correlation Error with Small Sample Sizes

Correlation error and average per capita RMS error did not substantially change even if sample sizes are small (see figures below). AR(1) continued to perform worse in capturing correlation. We estimated each method on a reduced number of sampled individuals, and we found that the RMS correlation error remained generally constant with an average of 0.07 and 0.14, respectively, for shortest distance and rank stability methods when the percentage sampled is above 50%. The RMS correlation error for the simulated values estimated using

AR(1) regression remained higher than shortest distance or rank stability for all sampled percentages (average of 0.15 when the percentage is over 50% and 0.31 over all sample percentages).

RMS Correlation Error in the Framingham Offspring Case Study With Smaller Numbers of Individuals

Per Capita RMS Error Per Time Period with Small Sample Sizes

Framingham Offspring per capita RMS error per time period, over the fraction of data sampled. All methods perform similarly, and shortest distance methods are not consistently the best- or worst-performing.

Autocorrelation Trends with Small Sample Sizes

Even when a smaller number of individuals are sampled, shortest distance methods do not show systematically higher autocorrelation than the other estimation methods. Here is the same autocorrelation table generated if the population consisted of only 10% of the original cohort.

Autocorrelation Table

Correlation Between Risk Factor Values at Time Period 1 and 4 in the Cohort, by Prediction Method, when only 10% of the cohort is used for the analysis

	True	Shortest Dist.	Shortest Dist.Wtd.	Rank Stability	AR(1) Reg.
BMI					
Mean	-0.135989	0.0212805	0.114837	0.0877083	0.0755688
SD	0.149508	0.247682	0.328823	0.41708	0.303184
HDL					
Mean	-0.127869	-0.0465573	0.0444516	0.0394001	0.0326162
SD	0.230363	0.213291	0.301746	0.491293	0.244624
TotChol					
Mean	0.066893	0.0359288	-0.0800401	0.0620588	-0.0733545
SD	0.192065	0.219476	0.269747	0.446901	0.28141
Glu					
Mean	-0.0753744	0.101487	0.138887	0.12962	0.15063
SD	0.135988	0.132812	0.221786	0.325628	0.239252
Tri					
Mean	0.159036	0.00219354	-0.102613	0.157111	0.0494585
SD	0.179276	0.268227	0.2152	0.392615	0.379514
BPrtt					
Mean	0.0205868	0.0368174	-0.0791722	0.0109341	0.0608945
SD	0.270915	0.270398	0.244103	0.373582	0.473792
Smoke					
Mean	0.0161669	-0.139657	-0.0154161	0.0432533	0.069814
SD	0.237712	0.278001	0.263255	0.262084	0.385424
SBP					
Mean	-0.0380902	0.204245	0.0345923	0.181638	0.0798771
SD	0.205813	0.228195	0.328258	0.463228	0.257883
DBP					
Mean	-0.0406508	0.18783	0.0208266	0.162539	0.0488966
SD	0.17332	0.292371	0.322947	0.440599	0.354542

Exploring Selective Mortality Effects

To explore the effect of mortality, we use a stylized example instead of incorporating risk-score dependent mortality models that may bring with them their own biases. In our stylized example, we assume that in our “true” data, all individuals with ASCVD risk scores above 30% die prior to being observed again. However, a simulation modeler would not be able to observe all the risk factors for the individuals who die; we assume he knows only the number of deaths in each period and that individuals with the higher ASCVD scores are more likely to die. We wish to test the simulation outcomes if such a modeler used each of the estimation methods we discuss. As in our base analysis, the more accurate the estimation method, the better match the ASCVD score should be with the true distribution.

To implement this example, the cross-sectional data for subsequent observations (for the shortest distance and rank stability methods) is drawn from the “true” population left alive, and the AR(1) coefficients are generated using data from individuals who survive the first two periods. We then implement all the prediction methods as before, with the additional change that the top individuals with the highest ASCVD scores in the simulation die, such that the number who die in the simulation match that of the “true” data.

In general, the results show that these methods, particularly shortest distance methods, can be reasonable choices even in the presence of selective mortality. The ASCVD distribution results are presented in the figure below. All methods were able to accurately capture the mean and interquartile range even in the presence of mortality. However, all methods tended to overestimate variation in ASCVD risk score, with rank stability and AR(1) methods showing a larger overestimation, particularly in the upper tail.

ASCVD 10-year Risk Score Distributions at the Last Time Period (T=4)

Exploring the Effect on Cost-Effectiveness Analyses

To illustrate how correlation accuracy might influence cost-effectiveness analysis (CEA) results, we performed a simplified example of a CEA where individuals in the Framingham Offspring Cohort case study were assigned an annual QALY weight proportional to their 10-year ASCVD risk scores such that individuals with 100% risk would be given a weight of 0.7 and individuals with 0% risk assigned a weight of 1 (perfect health). Similarly, in this simplified example CEA, medical costs were assumed to be proportional to ASCVD risk, with \$10k for those at 100% risk and \$0 for 0% risk. The 100% risk QALY weights and costs were taken from ranges reported in Pandya et al.

Total cohort QALYs and Costs over the four-period time horizon are shown in the figures below. The shortest distance and rank stability prediction methods are closer to the true QALY and cost values, most noticeably at the last time period. This should be expected given the poorer ASCVD risk score prediction by AR(1) regression, as seen in the Kolmogorov-Smirnov statistic results.

Plots of example QALYs and costs each time period (see table for means and SD over the full time horizon, over 10000 runs)

Method	Total Discounted QALYs		Total Discounted Costs	
	Mean	(5th, 95th pctl)	Mean	(5th, 95th pctl)
True	656.40	(655.36, 657.41)	1,324,320	(1290650, 1358890)
Short. Dist.	656.50	(655.45, 657.52)	1,320,990	(1287050, 1356010)
Short. Dist. Wtd	656.50	(655.45, 657.52)	1,320,990	(1286960, 1356210)
Rank Stability	656.53	(655.44, 657.58)	1,319,920	(1285070, 1356540)
AR(1)	656.00	(654.27, 657.68)	1,337,600	(1281740, 1395490)

In this example, we find that AR(1) generates the largest differences in total discounted QALYs and Costs from the true estimates (see table below), with about three times as much deviation from the true value as the other methods for Costs.

Method	Difference from True Total Discounted	
	QALYs	Costs
Short. Dist.	0.10	-3,330
Short. Dist. Wtd	0.10	-3,330
Rank Stability	0.13	-4,400
AR(1)	-0.40	13,280

The differences in QALY and cost estimations between these methods could influence whether an intervention qualifies as cost effective. In this example, the estimation methods overestimated the total discounted QALYs and underestimated the costs, but that does not always have to be the case. In general, these methods could under- or over-estimate the cost-effectiveness of an intervention, depending on how the intervention and base case were simulated. However, if the health outcomes or costs depend on a combination of risk factors, as ASCVD risk does, better capturing the correlation between risk scores may result in more accurate modeling and intervention prediction.

Shortest Distance Algorithm Code

Versions of Munkres algorithm are available for download for a variety of software platforms:

- In R [16]: https://artax.karlin.mff.cuni.cz/r-help/library/clue/html/solve_LSAP.html.
- In python [17]: <https://pypi.python.org/pypi/munkres/>
- In Matlab [18]: <http://www.mathworks.com/matlabcentral/fileexchange/20652-hungarian-algorithm-for-linear-assignment-problems--v2-3->

Utilization of Munkre's algorithm for risk factor value estimation

We used the following Matlab function to estimate the shortest distance method.

```

%%% Shortest distance method
function newStateMat = transitions_shortestDist2(K, stateMat, sampledPpl,
distanceWeights)
% INPUTS
% input is a N by m matrix of N number of people, m number of risk factors
% nextMean is a m x 1 vector of the risk factor means at the next period
% nextSigma is a m x m variance-covariance matrix of risk factors at the
next
period

```

```

%
% OUTPUTS
% newRF is a N by m matrix of the updated individual characteristics
% initialize
input = stateMat(:,K.RFcolumns);
numPpl = size(input,1);
numRFs = size(input,2);
newRF = -1*ones(numPpl, numRFs);
% Begin creating the estimated sample for the next time period
% Create standardized inputs for Munkres
nextSample = sampledPpl(:,K.RFcolumns);
demeanedNewSample = bsxfun(@minus,nextSample,mean(nextSample));
stdizedSample = bsxfun(@rdivide,demeanedNewSample,std(nextSample));
stdizedSample(isnan(stdizedSample)) = 0;
demeanedInput = bsxfun(@minus,input,mean(input));
stdizedInput = bsxfun(@rdivide,demeanedInput,std(input));
stdizedInput(isnan(stdizedInput)) = 0;
% Get distances to make a preference matrix
if distanceWeights == 0 %no distance weights
    distances = pdist2(stdizedInput,stdizedSample); %using matlab function
else
    %Compute a distance that weights each coordinate contribution differently
    weuc = @(XI,XJ,W)(sqrt(bsxfun(@minus,XI,XJ).^2 * W'));
    distances = pdist2(stdizedInput,stdizedSample, @(Xi,Xj)
weuc(Xi,Xj,distanceWeights));
end
% Implement Munkres algorithm
[assignment,~] = munkres(distances);
match = assignment';
%translate into stateMat variables
newStateMat = nextSample(match,:);
end

```

MATLAB Code to Generate Synthetic Data

This code creates synthetic longitudinal data (since longitudinal data was used in the Framingham Offspring case study). Similar to the Framingham Offspring case study, we sample individuals from each period to create cross sectional data that we can then run our estimation methods on and compare the outcomes to the true longitudinal data generated here.

MATLAB code used to generate synthetic data with increasing correlation over time:

```

% Randomly assign risk factor values by sampling from the normal

```

```

% distribution (mean 10, SD 4) for N people for T time periods to
% generate a "true" "longitudinal" dataset where rows are
% individuals, columns are risk factor values, and the 3rd
% dimension is time period.
%generate values for risk factor 1
longData(:,1,:) = repmat(normrnd(10,4,N,1),1,T) + repmat((1:T),N,1);
%generate values for risk factor 2
longData(:,2,:) = repmat(normrnd(10,4,N,1),1,T) + repmat((1:T),N,1);
%induce an increasing increasing correlation between RF1 and RF2
for tVal = 2:10
    somePpl = randsample(1:N, floor(N*tVal/10), false);
    longData(somePpl,noiseOffsetVal+varNumIdx,tVal) =
longData(somePpl,1,tVal);
end
%generate values for risk factor 3
longData(:,3,:) = normrnd(40,10,N,T);
%generate values for risk factor 4
longData(:,4,:) = normrnd(40,10,N,T);

```

MATLAB code used to generate oscillation synthetic data:

```

%%%% for synthetic data case study with oscillations
% make oscillations in 3 and 4
longData(:,3,1) = normrnd(10,4,N,1);
longData(:,4,1) = normrnd(40,10,N,1);
for t = 2:T
    longData(:,3,t) = longData(:,4,t-1);
    longData(:,4,t) = longData(:,3,t-1);
end
if onlyRF4oscillates == 1 %for scenario where only RF4 oscillates
    longData(:,3,1) = normrnd(10,4,N,1); %make RF3 random normal
end

```

Appendix References

19. Pandya A, Sy S, Cho S, Weinstein MC, Gaziano TA. Cost-effectiveness of 10-Year Risk Thresholds for Initiation of Statin Therapy for Primary Prevention of Cardiovascular Disease. JAMA. 2015; 314(2):142–150. Accessed online at: <http://jamanetwork.com/journals/jama/fullarticle/2396476>. [PubMed: 26172894]

Figure 1.
Schematic of Methods Using Example with 2 Risk Factors (BMI and HDL).

Figure 2.

Correlations and ASCVD Risk Estimates Across Methods at the Last Time Period. We provide a visualization of the pairwise Pearson correlation between risk factors estimated using each of the methods (a). The bars provide the mean correlation and the whiskers show the 5th–95th percentile range over the 10,000 iterations. Notably, the AR(1) regression method estimates differ significantly from the true correlations. Error in correlation can lead to differences in risk scores that depend on multiple risk factors, such as ASCVD. Panel b shows each method's accuracy in predicting high- and low-risk (as defined by a ASCVD cutoff score). AR(1) performs with lower accuracy for ASCVD cutoff values over 0.25. All pairwise correlations and false positive and true positive rates are shown in the Appendix.

Figure 3. Rank Stability When Correlation Increases Over Time. The pairwise Pearson correlation between risk factor 1 and 2 increases over time (a), but estimates generated using rank stability maintain the same correlation across all time periods (b).

Figure 4.

Performance with Large Fluctuations. Wide risk factor fluctuations (a) may influence the performance of all the methods (b).

Summary Statistics of Variables in Framingham Offspring Case Study (White Males Aged 55–65 in First Observation)

Table 1

Variable (Abbreviation)	Units	Average Values (5 th , 95 th percentile)			
		Observation 1 1979 – 1983	Observation 2 1983 – 1987	Observation 3 1987 – 1991	Observation 4 1991 – 1995
Age	Years	58 (55,63)	62 (59,67)	66 (63,71)	70 (66,74)
Body Mass Index (BMI)	kg/m ²	27 (22,33)	28 (23,35)	29 (23,35)	29 (23,35)
High Density Lipoprotein (HDL)	mg/dL	43 (28,63)	45 (29,67)	43 (28,64)	43 (26,66)
Total Cholesterol (TotChol)	mg/dL	218 (168,289)	223 (171,297)	208 (159,268)	200 (152,259)
Fasting Glucose (Glu)	mg/dL	106 (87,132)	104 (81,144)	102 (79,160)	110 (85,174)
Triglycerides (Tri)	mg/dL	150 (45,339)	160 (48,404)	156 (52,368)	161 (63,329)
Taking Blood Pressure Medication (BPMed)	Yes/No	0.23 (0,1)	0.35 (0,1)	0.41 (0,1)	0.41 (0,1)
Smoking Status (Smoke)	Yes/No	0.34 (0,1)	0.18 (0,1)	0.17 (0,1)	0.11 (0,1)
Systolic Blood Pressure (SBP)	mmHg	134 (109,168)	134 (110,163)	139 (112,172)	136 (110,173)
Diastolic Blood Pressure (DBP)	mmHg	83 (69,99)	81 (67,96)	80 (64,99)	73 (59,91)

Kolmogorov-Smirnov Statistic Results. Shortest distance methods and rank stability mostly resulted in distributions indistinguishable from the true distribution at the 5% significance level, while the ASCVD distribution estimated using AR(1) is much more likely to be statistically distinct from the true distribution across the 10,000 sampling iterations.

Table 2

Average and s.e. (in parentheses) values over 10,000 runs						
	Shortest Dist		Shortest Dist Wth		Rank Stability	AR(1)
Kolmogorov-Smirnov Statistic	0.076	(0.017)	0.076	(0.017)	0.088 (0.021)	0.105 (0.026)
P-Value	0.582	(0.234)	0.583	(0.235)	0.435 (0.254)	0.275 (0.225)
Fraction significant at the 5% level	0.001		0.002		0.026	0.145

Average per Capita RMS Error per Time Period for Each Risk Factor, reported in standard deviations. Error values have been normalized by the standard deviation for the appropriate risk factor at T=4. RMS error values are less than one standard deviation for many risk factors, and all are less than 1.5 standard deviations for all estimation methods.

Table 3

	BMI	HDL	TotChol	Glu	Tri	BPrrt	Smoke	SBP	DBP
Shortest Distance	0.79	0.97	1.02	0.78	1.13	0.92	1.08	0.99	1.01
Shortest Dist., Wtd	0.63	1.02	1.29	1.14	1.37	0.95	1.07	1.01	1.04
Rank Stability	0.54	0.79	0.89	0.74	1.01	1.01	1.04	0.85	0.90
AR(1) Regression	0.80	0.97	1.38	0.79	1.23	0.76	1.03	0.94	1.12