HHS Public Access

Author manuscript

IEEE Int Conf Connect Health Appl Syst Eng Technol. Author manuscript; available in PMC 2017 October 11.

Published in final edited form as:

IEEE Int Conf Connect Health Appl Syst Eng Technol. 2016 June ; 2016: 219–228. doi:10.1109/CHASE.2016.19

Multiple- vs Non- or Single-Imputation based Fuzzy Clustering for Incomplete Longitudinal Behavioral Intervention Data

Zhaoyang Zhang and **Hua Fang**

Division of Biostatistics and Health Services Research, Department of Quantitative Health Science, University of Massachusetts Medical School, Worcester, MA 01655

Abstract

Disentangling patients' behavioral variations is a critical step for better understanding an intervention's effects on individual outcomes. Missing data commonly exist in longitudinal behavioral intervention studies. Multiple imputation (MI) has been well studied for missing data analyses in the statistical field, however, has not yet been scrutinized for clustering or unsupervised learning, which are important techniques for explaining the heterogeneity of treatment effects. Built upon previous work on MI fuzzy clustering, this paper theoretically, empirically and numerically demonstrate how MI-based approach can reduce the uncertainty of clustering accuracy in comparison to non-and single-imputation based clustering approach. This paper advances our understanding of the utility and strength of multiple-imputation (MI) based fuzzy clustering approach to processing incomplete longitudinal behavioral intervention data.

Index Terms

longitudinal data; Missing values; Fuzzy clustering; Multiple imputation; MIFuzzy

I. Introduction

Behavioral interventions (e.g., physician advice; individual, group, or telephone counseling; self-help, including Internet use, and emerging use of wearable biosensors) [1]–[16] are commonly used to facilitate substance use treatments. Many such interventions have multiple components and are implemented over time, e.g., [17]–[32]. Over the course of an intervention, patients display complex and varying behaviors such as relapsing or dropping out for various psychological, social and environmental reasons, e.g., [33]–[41]. Their variations in engaging with or responding to interventions may contribute to different outcomes. Failure to appreciate these variations within, not only those between treatment and control groups, can ultimately lead to inappropriate and ineffective interventions, e.g., [38]–[42]. Thus, disentangling patients' behavioral variations, is a critical step for better understanding an intervention's effects on individual outcomes. However, capturing these variations poses significant computational and analytical challenges. Methods have been called for to address the variations in intervention effects for improving the design, conduct, and analyses of patient-oriented research [43]. Missing data also commonly exist in such complex longitudinal data, aggravating these methodological challenges. For these reasons, we propose an innovative trajectory pattern-recognition approach to tackle incomplete longitudinal behavioral intervention data, built upon our methods studied for observational

studies and our previous work on fuzzy clustering and missing data in longitudinal studies [44]–[52].

Multiple imputation (MI) has been well studied for statistical analyses [53], [54] but has not yet been scrutinized for clustering [55], [56]. One of major reasons is that we do not know how much the uncertainty of imputed data will affect the clustering accuracy. To answer this question, we need to compare the MI approach with single- (SI) and non-imputation (NI) based clustering approaches, which are commonly adopted in current pattern recognition areas. Although no single clustering approach can handle every type of data problem, our proposed MI-based Fuzzy clustering [48] was especially designed for longitudinal behavior related data that are typically non-normal, high dimensional, contain missing values, and of different types (i.e., continuous, ordinal and nominal). MI-Fuzzy was developed primarily because no other clustering technique overcomes all eight disadvantages for processing such data: 1) Cannot handle missing data directly; relies on list-wise/pair-wise deletion or single imputation; 2) Lack (available, tractable, replicable, or easy-to-use) validation indices or integrated validating processes; 3) Require statistical or prior distribution assumptions; 4) Require (manually) complicated parameter settings and repeated adjustment of model constraints and starting values of parameters in the clustering process; 5) Computationally intensive or malfunction for high-dimensional and longitudinal data with relatively large portions of missing values or increased number of clusters; 6) Results are (sometimes) not replicable (even using the same data set); 7) No graphic visualization of clusters from high-dimensional data or longitudinal data trajectories; 8) Unclear utility in behavior-related data.

For example, probabilistic clustering, including Gaussian Mixture models [57] and Hidden Markov Model-based Bayesian clustering [58]–[61] are commonly used for clustering. For such models, we must specify underlying statistical distributions (Gaussian) or prior distributions (Bayesian approach). The expectation-maximum (EM) algorithm for the Gaussian model is iterative and requires specifying initial values of parameters and adjusting parameters during modeling. EM often converges to a local maximum or does not converge at all; also, it can be computationally very slow when there are multi-modal distributions or mixture of Gaussians (e.g., potential clusters) and large proportions, say greater than 20%, of missing values with high-dimensional data and many clusters. Bayesian clustering is also computationally intensive and requires parameter adjustment and setting model constraints. Bayesian clustering has been developed and tested for large datasets (e.g., gene data) but its utility beyond gene analyses is unclear. If the observed data are generated by the assumed distributions, these probabilistic clustering models should be robust. However, high-dimensional behavioral intervention data typically have multi-modal distributions and their exact (joint) distributions are a priori unknown.

Neural networks models [62], [63] are mostly used for supervised learning (e.g., classification where the group labels of each subject are previously known). The well-known neural networks model for unsupervised learning (clustering where the group labels of each subject are a priori unknown) is called Kohonen's Self Organizing Map (SOM). The SOM architecture consists of an input layer and a Kohonen layer. The number of neurons (nodes) in the input layer corresponds to the number of attributes. Each input to the Kohonen layer is modified by a weight, which multiplies with the input value. SOM initializes weights by

assigning them random values. The Kohonen layer represents potential clusters. Hidden layers can also be added between the input and Kohonen layers to fit nonlinear data; however the number of neurons on this layer is arbitrary, and different settings lead to different results even for the same data. SOM has been applied in areas such as imaging, speech and voice recognition. Its validation indices or procedures are unclear. SOM itself cannot handle missing values. Its utility in behavioral intervention studies is unknown.

Hierarchical clustering [62], [64] has two major forms: Agglomerative (bottom-up) and Divisive (top-down). Agglomerative clustering starts with an every individual in his own cluster; at every step it merges the closest pair of clusters. Divisive clustering starts with all subjects in one cluster and split this cluster into smaller pieces. We will focus on agglomerative, as divisive clustering is much less used in applications [62], [64]. No matter what form, the disadvantages of hierarchical clustering are well-known: its inability to incorporate information about the shape and size of clusters, and its static nature (data are committed to a cluster once and cannot be moved to another cluster later [40]). It assumes the data have a hierarchical or nested structure which is implausible for, say, distinct components of behavioral interventions. Although it works in some areas, this model performed poorly in our preliminary prenatal tobacco exposure studies, yielding unreliable and trivial clusters (e.g., many small clusters with few individuals in each). With a pre-specified number of clusters, its clustering accuracy and inconsistency rates were the poorest in comparison to MI-Fuzzy and K-means. Partition-based clustering [62], [64] also comes in two flavors: hard (crisp) clustering that partitions the data set into mutually exclusive subsets, and fuzzy clustering that allows subjects to belong to several subsets but with different degrees of membership (see Section Innovation). The best-known hard clustering model is K-Means. It was designed to cluster numerical data into clusters in which each cluster has a center called mean. Although K-Means has been widely applied with appealing computational efficiency, it cannot handle missing values. Our preliminary analyses indicate that it has lower accuracy and higher inconsistency rates than MI-based fuzzy-clustering for pregnant smoking behavioral data.

Based on our prior research for longitudinal behavior intervention data, we will theoretically comparing MI- to SI- and NI-based fuzzy clustering approaches. In addition to our theoretical demonstration, we will use two real datasets. Table I are the notations of symbols used in this paper.

The rest of this paper is organized as follows: Section II introduces MI-based Fuzzy clustering. Section III theoretically compares and demonstrates MI-, SI- and NI-based fuzzy clustering approaches. Section IV presents numerical results under three missing data mechanisms. Section V concludes the paper.

II. Multiple Imputation Based Fuzzy Clustering

To compare the three missing data imputation approaches, two longitudinal behavioral intervention datasets will be used as examples to demonstrate our MI-based approach: TDTA [65], [66] and QuitPrimo (QP) [67]. TDTA data were collected from a culturally-adapted smoking cessation intervention for Korean Americans. Intervention attributes in MI-

Fuzzy were based on data availability and/or selected to maximize information that depicts individual response variations resulting from their psychological factors and engagement with the intervention. All designed intervention components and time for cessation counseling were included at the beginning. The three components were (a) cognitive behavioral therapy, (b) cultural adaptation, and (c) nicotine replacement therapy. The first two components described smokers' psychological reactions to the culturally adapted cognitive behavioral therapy, which were measured by scores on Perceived Risks and Benefits of Quitting Smoking, Perceived Family and Peer Norm for Quitting, and Self-efficacy in Quitting scales. Each scale has four repeated measures collected at four follow-ups: 1, 3, 6, and 12 months from the quit day, total 20 intervention attributes (5 subscales×4 times). The last component "Nicotine Replacement Therapy" was measured by the number of nicotine patches returned after use (1 attribute) and the counseling time was measured in the unit of minutes (1 attribute). In all, 22 intervention attributes were initially used for MI-Fuzzy clustering. The percentage of missingness of TDTA data ranges from 9% to 18% on each of tested intervention attribute; and 65% participants have all values, 34% have more than 5 values on intervention attributes.

QP is a two-arm longitudinal web trial used to assist in the smoking cessation of a general smoker population. The data were collected via an online referral portal with 1320 individuals. The intervention arm was engaged with three components which the control arm cannot see. The first intervention component describes how often smokers communicate with a tobacco treatment specialist in a secure form. The second one measures how often the smokers are engaged or encouraged by experts. The third main component describes how many times smokers view messages and dialogue from peers and ex-smokers through a resource website. Each smoker has six monthly measures for each component for six months (total 18 attributes).

MI-Fuzzy is the first clustering model to date that employs a full theoretical integration of (a) multiple imputation (MI), (b) fuzzy clustering, and (c) comprehensive validation [48], [49]. It simultaneously copes with real-world situations where patients have membership in multiple clusters, handles high-dimensional longitudinal intervention data with missing values (e.g. multiple repeatedly-measured correlated constructs), and validates response patterns.

Our MI process was described in detail elsewhere and implemented in Matlab [68]. Briefly, we assume intervention data follow an arbitrary missing pattern, where other missing patterns such as monotone patterns are special cases [69]. Imputed datasets were generated using the Markov Chain Monte Carlo (MCMC) method with multiple chains, non-informative Jeffreys prior of the Bayesian approach, and 500 burn-in iterations. Before the MCMC process, an expectation-maximization (EM) procedure is conducted to train the parameters of the data. At each MCMC iteration, an I-step is conducted to fill the data by random draws from the model according to the trained parameters, followed by a P-step to update the parameters in the new completed data. The I-step and P-step are conducted iteratively until the Markov Chains stabilized.

The fuzzy clustering aims to minimize an objective function

$$f_m(X,U,V)=\sum_{k=1}^{c}\sum_{i=1}^{n}(u_{ik})^w\|x_i-v_k\|^2$$
$$\text{subject to } \sum_{k=1}^{c}u_{ik}=1,\forall i, \quad (1)$$

where X is the dataset, V is the cluster centroids, k is the k-th cluster, w is parameter of U is a vector of u_{ik} where $0 \le u_{ik} \le 1$, $\forall$ i, k denotes the fuzzy degree of membership for subjects i, ($i = 1, 2, \ldots, n$) in the respective cluster k. In MI-Fuzzy, we minimized this fuzzy objective function (e.g., we minimized the intra-cluster variance) for each imputed data set,

$$f_m(X,U,V)=\sum_{k=1}^{c}\sum_{i=1}^{n}(u_{ik})^w\|x_i-v_k\|^2+\sum_{i=1}^{n}\lambda\left(\sum_{k=1}^{c}u_{ik}-1\right), \quad (2)$$

where λ is a weight and $\lambda \ge 0$.

Intervention data have a unique feature that can be used to evaluate clustering accuracy. For example, MI-Fuzzy can be evaluated based on the intervention clustering accuracy rate (ℑ) for treatment data where we know the labels with certainty (for example, who is in the control vs. intervention group), but not whether there will be different clusters (subgroups) within the intervention group due to behavioral variations. This rate is calculated as the average clustering accuracy rates across M imputed datasets:

$$\Im=\frac{1}{M}\sum_{i=1}^{M}(N-n_{control}-n_{intervention})/N, \quad (3)$$

where N is the total sample size, $n_{control}$ is the number of mislabeled control group members and $n_{intervention}$ is the number of mislabeled intervention group members. The larger the rate the more accurate the clustering.

Given a termination clustering number (CT), $CT=\sqrt{N/2}$, where N is the sample size, the MI-Fuzzy algorithm searched for the optimal number of clusters through a comprehensive validation procedure [48], [49], [51].

III. MI- VS. SI- AND NI- FUZZY CLUSTERING

For statistical analyses, MI-based parameter estimation were shown more reliable than single imputation approach (e.g., mean, regression, and hot deck) which can introduce bias or lose precision. However, for clustering, MI-based approach has not yet been theoretically demonstrated better than other imputations, although intuitively would improve the uncertainty of clustering accuracy. Below we showed our theoretical demonstration.

1) MI vs Non-imputation (NI) Clustering

Lemma 1. MI- achieves higher clustering accuracy rate than NI-based Fuzzy Clustering

Proof: Let X denote the $N * d$ data matrix, where n denotes the number of patients who belong to k clusters and d the number of attributes. Let X_{mis} denote the missing part of X, and X_{obs} denotes the observed part with N' patients, $N' < N$, who also belong to these k clusters. Assume the n patients are clustered into the k clusters with accuracy $\Im$. With NI-fuzzy clustering, $(N - N')$ patients are not clustered due to their missing values on d. Then the clustering accuracy for NI is

$$\Im_{NI} = \Im \times N'/N, \quad (4)$$

For MI, X_{mis} are multiply-imputed, therefore, N patients are clustered. Suppose N^{MI} $(0 < N^{MI} \leq N - N')$ patients with X_{mis} is correctly clustered, then the accuracy is,

$$\Im_{MI} = \Im(N' + N^{MI})/N. \quad (5)$$

Therefore, $\Im_{MI} > \Im_{NI}$. Thus, the MI-based achieves higher accuracy than NI-based fuzzy clustering.

2) MI vs Single Imputation (SI) Clustering

The most commonly used single imputation methods include mean, regression and hot-deck imputation [70], which are used for demonstration in comparison to MI based clustering hereafter.

Lemma 2. MI-based approach reduces the uncertainty of clustering accuracy for incomplete longitudinal data compared to SI-based fuzzy clustering

Proof

1) Mean based SI: The missing values are filled by the means of the observed values. The variance of data is underestimated after mean imputation. For clustering, the patients close to each other are more likely to be classified into the same cluster. Since the missing values are filled by the same value (the mean), the patients with these means become closer to each other, and likely to be classified into the same cluster.

For a completed d-dimensional data X with n patients, suppose X contains k clusters where each cluster has N/k patients. Let r be the percentage of missingness, then each cluster has rN/k patients with missing values. The missing values are imputed by the mean SI before clustering. Then patients with imputed values (μ) are likely to be clustered in the same cluster and the mean SI-based clustering accuracy is expressed as

$$\Im_{SI-mean}=1-rN(k-1)/k, \quad (6)$$

where $rN(k-1)/k$ is the number of mistakenly clustered patients.

Figure 1a demonstrates a scenario where the mean-imputation reduces the clustering accuracy $\Im$. The scatter plot shows a two-dimentional X represented by (x_{i1}, x_{i2}), where k = 2. The missing values of x_{i1} are marked by green circles in the 1st cluster and by green squares in the 2nd cluster. The patient i with missing values are denoted by

$$X_i-mis=\{(x_{i1-mis}, x_{i2})\}. \quad (7)$$

By conducting one-time mean-imputation, all the missing values are filled by $x_{1\mu}$ (marked by red circles/squares), then the patients are represented by

$$X_{SI-mean}=\{(x_{i\mu}, x_{i2})\}, \quad (8)$$

where $x_{i\mu}=\frac{1}{n'}\sum_{i=1}^{n'}x_{i1}$ and n' is the number of observed x_{i1} from all patients.

This one-time imputed values are the same and treated as if they are observed values and then used for clustering. As demonstrated in Figure 1b, 50% of the patients with missing values can be mistakenly clustered into the 1st cluster marked by circles.

2) Regression based SI clustering: A regression model can be learned from X_{obs} and the learned model with estimated coefficient is used to impute missing values of a variable based on other variables. For illustrative purposes, the regression co-efficient $\hat{\beta}$ can be learned based on observed two-dimensional data, e.g.,

$$x_{i2}=\hat{\beta}x_{i1}+\varepsilon, \quad (9)$$

then

$$x_{i2-mis}=\hat{\beta}x_{i1}. \quad (10)$$

Figure 2a demonstrates a scenario $\Im_{SI\text{-}reg}$ where the regression SI reduces the clustering accuracy. With the same notation, the patient i with missing values on the second variable are denoted by

$$X_{i-mis}=\{(x_{i1}, x_{i2-mis})\}. \quad (11)$$

The missing values are imputed only once by this regression SI (marked by red circles/squares), and more likely to fit perfectly along the regression line since the imputed data do not have an error term included in their estimation,

$$X_{SI-regression}=\left\{(x_{i1}, \hat{\beta}x_{i1})\right\}. \quad (12)$$

Again, this one-time imputed values with errors are treated as if they are actual observed values and used for clustering. In the worst case, all cases with missing values can be mistakenly clustered, as shown in Figure 2b.

3) Hot-deck SI clustering: This type of SI first sorts a dataset according to any variables, and fill the missing values by using the observed values immediately prior to them and repeat this process until all missing values are imputed. If the selected variable that used to order the data does not have a significant role in clustering patients, the imputed data could be fully messed up after sorting and the likelihood of clustering the patients with missing values to their corresponding clusters is small.

Figure 3a shows how hot-deck SI reduces the clustering accuracy. Similarly, for a two-dimensional data X, represented by ($x1$, $x2$), where $k = 2$. The missing values of x_1 are marked by green or circles squares in these two clusters, respectively. By conducting the hot-deck SI, the data are sorted by x_1, and the missing values are filled by the observed x_1 values just prior to them (marked by red circles/squares),

$$X_{SI-hotdeck}=\{(x_{i1}, x_{j2})\}, \quad (13)$$

where j is the index of the case just before the i-th cases ordered according to x_2. As shown in Figure 3b, the cases with missing values could be completely misplaced in different clusters using this imputation.

Overall, the $\Im_{SI}$ can be computed as,

$$\Im_{SI}=(N-n_{control}-n_{intervention})/N. \quad (14)$$

Although SI-derived data have the appearance of completeness, it can mislead the clustering results because SI again does not account for the imputation uncertainty and clustering accuracy from such data is uncertain.

4) MI-clustering: Unlike SI methods, the uncertainty of clustering accuracy is accounted by computing the average of clustering accuracy from clustering each of multiply imputed datasets generated from the stochastic processes. The variance of $\mathfrak{J}_{MI}$ aforementioned above is computed as,

$$\mathrm{var}\left(\mathfrak{J}_{MI}\right)=\frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\mathfrak{J}}-\overline{\mathfrak{J}}\right)^{2}, \quad (15)$$

where M is the number of imputations which can be estimated given the expected relative efficiency and the missing information of parameter estimates of attributes [71]. As shown in Table II, given our expected relative efficiency of 0.98 and missing information for TDTA (20%) and QP (10%), we chose 10 imputations for each.

The confidence interval of $\mathfrak{J}_{MI}$ could be estimated as:

$$\mathfrak{J}_{MI}\pm 1.96\times\sqrt{\mathrm{var}(\mathfrak{J}_{MI})/N}. \quad (16)$$

For example, given two dimensions, x_1 and x_2,

$$\begin{array}{cc} x_1 & x_2 \\ \hline 3 & 6 \\ 5 & \cdot \\ 6 & 12 \end{array}$$

and assume the true value for the missing value is 4, the mean SI would fill the missing value as 9, regression SI will estimate 10 ($x_2 = 2 \times x_1$), and hotdeck sorts x_2 and gives 12. Then these values will treated as “true”. Differently, MI can estimate the missing value 10 times to account for the imputation uncertainty and thus, for clustering, it can reduce the uncertainty of clustering accuracy. As shown in Figure 4a, two patients marked green in each cluster have missing values on x_1 or x_2. Assuming missing values are imputed 10 times, based on the clustering results from the ten imputed datasets, the two patients would still stay in their own clusters, even if the ten impute values can differ widely, as shown in Figure 4(b). Thus, the imputation uncertainty is accounted in MI process and the consequent uncertainty of $\mathfrak{J}_{MI}$ is reduced.

Overall, our theoretical demonstration indicates MI-based clustering is superior to NI- and SI-clustering because it accounts for the uncertainty of imputation and clustering accuracy for incomplete data.

IV. NUMERICAL ANALYSES

This section evaluates the performance of MI-based fuzzy clustering in comparison to SI- and NI-clustering using TDTA and QP data. In addition, we simulated data using the

parameters from these two behavioral intervention studies. Specifically, we evaluated these three imputation-based clustering approaches under three missing data mechanisms. We also varied the percentage of missingness to evaluate the clustering performance based on the same validation index, called overlap over septation validation (OOS [51]) and clustering accuracy.

A. The Missing Mechanisms

In statistical theories, the three missing data mechanisms are termed as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [72], [73]. Using incomplete TDTA and QP data as an example, let X be the complete data without missing values, $X = X_{obs} \cup X_{mis}$, X_{mis} are the missing values on observed attributes P in the incomplete data X; and P' are unobserved attributes. For example, for QP web trial study, we have longitudinal zero-inflated count data, then zero-inflated Poisson (ZIP) models are considered for data simulation under the generalized linear mixed model (GLMM) framework while the typical GLMM were used to generate data using parameters from TDTA.

1. *Missing Complete At Random (MCAR):* we simulated data assuming X_{mis} are missing independently of both observed attributes P and unobserved P'.
2. *Missing At Random (MAR):* Data were simulated assuming X_{mis} are missing independently of unobserved P'.
3. *Missing Not At Random (MNAR):* Data were simulated assuming X_{mis} relate to unobserved attributes P'. MNAR occurs when the condition of MAR is violated.

B. Performance Evaluation Using TDTA and QuitPrimo

As shown in Figure 5, for TDTA, MI-based fuzzy clustering identified the optimal number of clusters as three, because the validation index reaches the lowest value, with clustering accuracy of 100%. SI-based clustering identified 3 clusters with accuracy of 85% and NI-based identified 3 clusters with accuracy of 80%.

For QP, MI Fuzzy identified 4 clusters with with clustering accuracy of 78%, as shown in Figure 6. SI-based clustering identified 5 clusters with accuracy of 70% and NI-based identified 4 clusters with accuracy of 68%.

As demonstrated in the theoretical section, SI- and NI-based approach do not account for the uncertainty of imputation and treated the imputed values as if they are observed for clustering. Therefore, their clustering results could be misleading, and the uncertainty of clustering accuracy is also not counted. Even if they identified the same clusters as MI-, the results could be due to chance.

We also evaluated the three imputation clustering approaches using simulation, where we primarily focus on their performance with a medium percentage of missingness (10%–15%) for each missing mechanism. We set the sample size n at 1000 and the dimension d at 20.

C. Performance under MCAR

As shown in Figure 7, for MCAR, MI-based fuzzy clustering identified the correct number of clusters, 4, and achieves 95% and 94% accuracy for 10% and 15% missingness, respectively. SI-based fuzzy clustering identified 4 clusters, with accuracy of 89% and 87% for 10% and 15% missingness, respectively. NI-based fuzzy clustering identified 4 clusters, with accuracy of 85% and 84% for 10% and 15% missingness, respectively.

D. Performance under MAR

As shown in Figure 8, for MAR, MI-based fuzzy clustering identified the correct number of clusters, 4, and achieves the accuracy of 96% and 93% for 10% and 15% missingness, respectively. SI-based fuzzy clustering identified 4 clusters, with accuracy of 88% and 85% for 10% and 15% missingness, respectively. NI-based fuzzy clustering identified 4 clusters, with accuracy of 86% and 83% for 10% and 15% missingness, respectively.

E. Performance under MNAR

As shown in Figure 9, for MNAR, MI-based fuzzy clustering identified the correct number of clusters, 4, and achieves the accuracy of 90% and 89% for 10% and 15% missingness, respectively. SI-based fuzzy clustering identified 4 clusters, with accuracy of 81% and 82% for 10% and 15% missingness, respectively. NI-based fuzzy clustering identified 4 clusters, with accuracy of 80% and 79% for 10% and 15% missingness, respectively.

Overall, under three mechanisms, MI-based clustering performance are similar and seems invariant to different missing data mechanisms. This invariant property of MIFuzzy may attribute to the fact that the actual MIFuzzy algorithm does not learn the ”concepts” of missing mechanisms but is somewhat sensitive to the real numbers such as the percentage of missingness. With the increase of the percentage of missingness, MI-based clustering accuracy decreased but still achieved at least 93% across the three mechanisms. Although SI and NI identified the same number of clusters, as discussed in the theoretical section, it is likely due to chance and their clustering accuracy is lower and misleading, as they cannot account for the imputation uncertainty and therefore the accuracy is uncertain.

V. Conclusion

Reliable and more accurate trajectory pattern recognition approaches will help capture patients’ variations in engaging with or responding to behavioral interventions. Explaining these observed variations within and between treatment and control groups can ultimately lead to appropriate and effective interventions adaptive to patients with at-risk patterns. This paper advances our understanding of the utility and strength of multiple-imputation (MI) based fuzzy clustering approach to processing incomplete longitudinal behavioral intervention data. Specifically, this paper theoretically, empirically and numerically demonstrated how MI-based approach can reduce the uncertainty of clustering accuracy in comparison to non-and single-imputation based clustering approach. Our future research will further evaluate MI-based approach under a large percentage of missingness and focus on visualization-aided validation to further expand the MI-based clustering approaches.

Acknowledgments

This research was supported by NIH grant R01 DA03332301, 1UL1RR031982-01 Pilot Project to Dr. Fang. We thank Dr. Thomas Huston and Dr. Sun Kim for providing their longitudinal TDTA data and web-delivered QuitPrimo trial data.

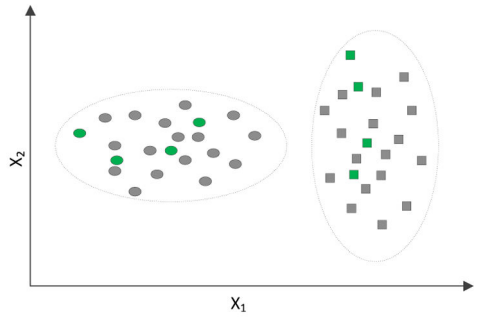
References

1. Sawicki P, Didjurgeit U, Mühlhauser I, Berger M. Behaviour therapy versus doctor's anti-smoking advice in diabetic patients. Journal of internal medicine. 1993; 234(4):407–409. [PubMed: 8409838]
2. Marlow SP, Stoller JK. Smoking cessation. Respiratory care. 2003; 48(12):1238–1256. [PubMed: 14651764]
3. Mottillo S, Filion KB, Bélisle P, Joseph L, Gervais A, O'Loughlin J, Paradis G, Pihl R, Pilote L, Rinfret S, et al. Behavioural interventions for smoking cessation: a meta-analysis of randomized controlled trials. European heart journal. 2009; 30(6):718–730. [PubMed: 19109354]
4. Barth J, Bengel J, Critchley J. Efficacy of psychosocial interventions for smoking cessation in patients with coronary heart disease: a systematic review and meta-analysis. Annals of Behavioral Medicine. 2006; 32(1):10–20. [PubMed: 16827625]
5. Hennrikus DJ, Lando HA, McCarty MC, Klevan D, Holtan N, Huebsch JA, Jestus S, Pentel PR, Pine D, Sullivan S, et al. The team project: the effectiveness of smoking cessation intervention with hospital patients. Preventive medicine. 2005; 40(3):249–258. [PubMed: 15533536]
6. Allen B Jr, Pederson LL, Leonard EH. Effectiveness of physicians-in-training counseling for smoking cessation in african americans. Journal of the National Medical Association. 1998; 90(10): 597. [PubMed: 9803724]
7. Hajek P, Taylor TZ, Mills P. Brief intervention during hospital admission to help patients to give up smoking after myocardial infarction and bypass surgery: randomised controlled trial. Bmj. 2002; 324(7329):87–89. [PubMed: 11786452]
8. Jackson AA, Manan WA, Gani AS, Eldridge S, Carter YH. Beliefs and behavior of deceivers in a randomized, controlled trial of anti-smoking advice at a primary care clinic in kelantan, malaysia. Southeast Asian journal of tropical medicine and public health. 2004; 35:748–755. [PubMed: 15689099]
9. Miller NH, Smith PM, DeBusk RF, Sobel DS, Taylor CB. Smoking cessation in hospitalized patients: results of a randomized trial. Archives of Internal Medicine. 1997; 157(4):409–415. [PubMed: 9046892]
10. Aveyard P, Griffin C, Lawrence T, Cheng K. A controlled trial of an expert system and self-help manual intervention based on the stages of change versus standard self-help materials in smoking cessation. Addiction. 2003; 98(3):345–345. [PubMed: 12603234]
11. Molyneux A, Lewis S, Leivers U, Anderton A, Antoniak M, Brackenridge A, Nilsson F, McNeill A, West R, Moxham J, et al. Clinical trial comparing nicotine replacement therapy (nrt) plus brief counselling, brief counselling alone, and minimal intervention on smoking cessation in hospital inpatients. Thorax. 2003; 58(6):484–488. [PubMed: 12775857]
12. Mogielnicki RP, Neslin S, Dulac J, Balestra D, Gillie E, Corson J. Tailored media can enhance the success of smoking cessation clinics. Journal of behavioral medicine. 1986; 9(2):141–161. [PubMed: 3712426]
13. Jorenby DE, Smith SS, Fiore MC, Hurt RD, Offord KP, Croghan IT, Hays JT, Lewis SF, Baker TB. Varying nicotine patch dose and type of smoking cessation counseling. Jama. 1995; 274(17):1347–1352. [PubMed: 7563558]
14. Lando HA, Rolnick S, Klevan D, Roski J, Cherney L, Lauger G. Telephone support as an adjunct to transdermal nicotine in smoking cessation. American Journal of Public Health. 1997; 87(10): 1670–1674. [PubMed: 9357351]
15. Carreiro S, Fang H, Zhang J, Wittbold K, Weng S, Mullins R, Smelson D, Boyer EW. imstrong: Deployment of a biosensor system to detect cocaine use. Journal of medical systems. 2015; 39(12):1–8. [PubMed: 25600193]

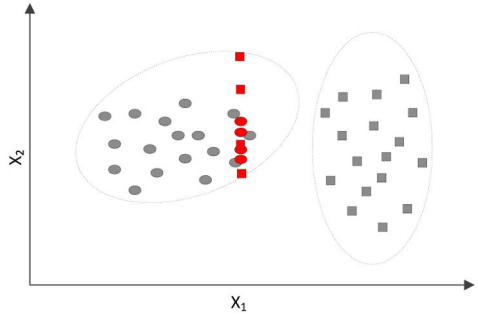
16. Fang, H., Wittbold, K., Weng, S., Stephanie, C., Zhang, J., Mullins, R., Boyer, E. ANNALS OF BEHAVIORAL MEDICINE. Vol. 49. SPRINGER; 233 SPRING ST, NEW YORK, NY 10013 USA: 2015. Describing real-time substance-use detection from big biosensor data: A case study of cocaine users; p. S91-S91.
17. Kim SS, Kim SH, Fang H, Kwon S, Shelley D, Ziedonis D. A culturally adapted smoking cessation intervention for korean americans: a mediating effect of perceived family norm toward quitting. Journal of Immigrant and Minority Health. 2015; 17(4):1120–1129. [PubMed: 24878686]
18. Fiore, MC., Bailey, WC., Cohen, SJ., Dorfman, SF., Goldstein, MG., Gritz, ER., Heyman, RB., Jaen, CR., Kottke, TE., Lando, HA., et al. Treating tobacco use and dependence: a clinical practice guideline. Publications Clearinghouse; 2000.
19. Fiore MC, McCarthy DE, Jackson TC, Zehner ME, Jorenby DE, Mielke M, Smith SS, Guiliani TA, Baker TB. Integrating smoking cessation treatment into primary care: an effectiveness study. Preventive medicine. 2004; 38(4):412–420. [PubMed: 15020174]
20. Glasgow RE, Lichtenstein E. Long-term effects of behavioral smoking cessation interventions. Behavior Therapy. 1987; 18(4):297–324.
21. Glasgow RE, Whitlock EP, Eakin EG, Lichtenstein E. A brief smoking cessation intervention for women in low-income planned parenthood clinics. American Journal of Public Health. 2000; 90(5):786. [PubMed: 10800431]
22. Secker-Walker RH, Solomon LJ, Flynn BS, Skelly JM, Lepage SS, Goodwin GD, Mead PB. Individualized smoking cessation counseling during prenatal and early postnatal care. American journal of obstetrics and gynecology. 1994; 171(5):1347–1355. [PubMed: 7977545]
23. Windsor RA, Woodby LL, Miller TM, Hardin JM, Crawford MA, DiClemente CC. Effectiveness of agency for health care policy and research clinical practice guideline and patient education methods for pregnant smokers in medicaid maternity care. American journal of obstetrics and gynecology. 2000; 182(1):68–75. [PubMed: 10649158]
24. Weissfeld JL, Holloway JL. Treatment for cigarette smoking in a department of veterans affairs outpatient clinic. Archives of internal medicine. 1991; 151(5):973–977. [PubMed: 2025147]
25. Slovinec D'Angelo ME, Reid RD, Hotz S, Irvine J, Segal RJ, Blanchard CM, Pipe A. Is stress management training a useful addition to physician advice and nicotine replacement therapy during smoking cessation in women? results of a randomized trial. American Journal of Health Promotion. 2005; 20(2):127–134. [PubMed: 16295704]
26. Omenn G, Thompson B, Sexton M, Hessol N, Breitenstein B, Curry S, Michnich M, Peterson A. A randomized comparison of worksite-sponsored smoking cessation programs. American journal of preventive medicine. 1987; 4(5):261–267.
27. Curry SJ, Marlatt GA, Gordon J, Baer JS. A comparison of alternative theoretical approaches to smoking cessation and relapse. Health Psychology. 1988; 7(6):545. [PubMed: 3215161]
28. Curry SJ, McBride CM. Relapse prevention for smoking cessation: review and evaluation of concepts and interventions. Annual review of public health. 1994; 15(1):345–366.
29. Glasgow RE, Schafer L, O'Neill HK. Self-help books and amount of therapist contact in smoking cessation programs. Journal of consulting and clinical psychology. 1981; 49(5):659. [PubMed: 7287975]
30. Rabius V, McAlister AL, Geiger A, Huang P, Todd R. Telephone counseling increases cessation rates among young adult smokers. Health Psychology. 2004; 23(5):539. [PubMed: 15367074]
31. Curry SJ, McBride C, Grothaus LC, Louie D, Wagner EH. A randomized trial of self-help materials, personalized feedback, and telephone counseling with nonvolunteer smokers. Journal of consulting and clinical psychology. 1995; 63(6):1005. [PubMed: 8543703]
32. Taylor CB, Houston-Miller N, Killen JD, DeBusk RF. Smoking cessation after acute myocardial infarction: effects of a nurse-managed intervention. Annals of internal medicine. 1990; 113(2): 118–123. [PubMed: 2360750]
33. Schoedel KA, Hoffmann EB, Rao Y, Sellers EM, Tyndale RF. Ethnic variation in cyp2a6 and association of genetically slow nicotine metabolism and smoking in adult caucasians. Pharmacogenetics and Genomics. 2004; 14(9):615–626.

34. Proudnikov D, Hamon S, Ott J, Kreek MJ. Association of polymorphisms in the melanocortin receptor type 2 (mc2r, acth receptor) gene with heroin addiction. Neuroscience letters. 2008; 435(3):234–239. [PubMed: 18359160]
35. Dagher A, Tannenbaum B, Hayashi T, Pruessner JC, McBride D. An acute psychosocial stress enhances the neural response to smoking cues. Brain research. 2009; 1293:40–48. [PubMed: 19632211]
36. Anderson KG, Ramo DE, Brown SA. Life stress, coping and comorbid youth: An examination of the stress-vulnerability model for substance relapse. Journal of Psychoactive Drugs. 2006; 38(3): 255–262. [PubMed: 17165368]
37. Brown SA, Vik PW, Patterson TL, Grant I, Schuckit MA. Stress, vulnerability and adult alcohol relapse. Journal of studies on alcohol. 1995; 56(5):538–545. [PubMed: 7475034]
38. Longshore D. Treatment motivation among mexican american drug-using arrestees. Hispanic Journal of Behavioral Sciences. 1997; 19(2):214–229.
39. Sabogal F, Otero-Sabogal E, Pasick R, Jenkins C, Perez-Stable E. Printed health education materials for diverse communities: suggestions learned from the field. Health Educ Q. 1996; 23
40. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Quarterly. 2004; 82(4):661–687. [PubMed: 15595946]
41. Baker TB, Mermelstein R, Collins LM, Piper ME, Jorenby DE, Smith SS, Christiansen BA, Schlam TR, Cook JW, Fiore MC. New methods for tobacco dependence treatment research. Annals of Behavioral Medicine. 2011; 41(2):192–207. [PubMed: 21128037]
42. Resnicow K, Soler R, Braithwaite RL, Ahluwalia JS, Butler J. Cultural sensitivity in substance use prevention. Journal of community psychology. 2000; 28(3):271–290.
43. Helfand M, Tunis S, Whitlock EP, Pauker SG, Basu A, Chilingerian J, Harrell FE Jr, Meltzer DO, Montori VM, Shepard DS, et al. A ctsa agenda to advance methods for comparative effectiveness research. Clinical and translational science. 2011; 4(3):188–198. [PubMed: 21707950]
44. Zhang Z, Fang H, Wang H. Visualization aided engagement pattern validation for big longitudinal web behavior intervention data. IEEE HealthCOM. 2015
45. Wang, CJ., Fang, H., Wang, C., Daneshmand, M., Wang, H. A novel initialization method for particle swarm optimization-based fcm in big biomedical data. proceedings of the 2015 IEEE International Conference on Big Data; IEEE; 2015. p. 2942-2944.
46. Wang, CJ., Fang, H., Wang, H. Dag-searched and density-based initial centroid location method for fuzzy clustering of big biomedical data. Proceedings of the 8th International Conference on Bioinspired Information and Communications Technologies; ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering); 2014. p. 290-293.
47. Wang CJ, Fang H, Kim S, Moormann A, Wang H. A new integrated fuzzifier evaluation and selection (nifes) algorithm for fuzzy clustering. Journal of Applied Mathematics and Physics. 2015; 3(07):802.
48. Fang H, Espy KA, Rizzo ML, Stopp C, Wiebe SA, Stroup WW. Pattern recognition of longitudinal trial data with nonignorable missingness: An empirical case study. International journal of information technology & decision making. 2009; 8(03):491–513. [PubMed: 20336179]
49. Fang H, Rizzo ML, Wang H, Espy KA, Wang Z. A new nonlinear classifier with a penalized signed fuzzy measure using effective genetic algorithm. Pattern recognition. 2010; 43(4):1393–1401. [PubMed: 20300543]
50. Zhang Z, Fang H. An enhanced visualization method to aid behavioral trajectory pattern recognition infrastructure for big longitudinal data. IEEE Transaction on Big Data. 2016 submitted.
51. Zhang Z, Fang H, Wang H. A new mi-based visualization aided validation index for mining big longitudinal web trial data. IEEE Access. 2016 accepted.
52. Wang, H., Fang, H., Sharif, H., Wang, Z. Nonlinear classification by genetic algorithm with signed fuzzy measure. Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International; IEEE; 2007. p. 1-6.
53. Rubin, DB. Multiple imputation for nonresponse in surveys. Vol. 81. John Wiley & Sons; 2004.
54. Schafer, JL. Analysis of incomplete multivariate data. CRC press; 1997.

55. Fang H, Johnson C, Stopp C, Espy KA. A new look at quantifying tobacco exposure during pregnancy using fuzzy clustering. Neurotoxicology and teratology. 2011; 33(1):155–165. [PubMed: 21256430]
56. Fang H, Dukic V, Pickett KE, Wakschlag L, Espy KA. Detecting graded exposure effects: A report on an east boston pregnancy cohort. Nicotine & Tobacco Research. 2012:ntr272.
57. McLachlan, G., Peel, D. Finite mixture models. John Wiley & Sons; 2004.
58. François O, Ancelet S, Guillot G. Bayesian clustering using hidden markov random fields in spatial population genetics. Genetics. 2006; 174(2):805–816. [PubMed: 16888334]
59. Zhang Z, Wang H, Lee KC, Fang H. Bridging vital signs and social interactions for resource-constrained epidemic control. International Journal of Information Technology & Decision Making. 2013; 12(03):469–489.
60. Wang, H., Fang, H., Xing, L., Chen, M. An integrated biometric-based security framework using wavelet-domain hmm in wireless body area networks (wban). Communications (ICC), 2011 IEEE International Conference on; IEEE; 2011. p. 1-5.
61. Wang, H., Fang, H., Espy, KA., Peng, D., Sharif, H. A bayesian multilevel modeling approach for data query in wireless sensor networks. Computational Science–ICCS; 2007; Springer; 2007. p. 859-866.
62. Gan G, Ma C, Wu J. Data clustering: theory, algorithms, and applications. Siam. 2007:20.
63. Kubat, M. Neural networks: a comprehensive foundation by simon haykin, macmillan, 1994. 1999. p. 409-412.
64. Keller, J., Krisnapuram, R., Pal, NR. Fuzzy models and algorithms for pattern recognition and image processing. Vol. 4. Springer Science & Business Media; 2005.
65. Kim SS, Kim S-H, Fang H, Kwon S, Shelley D, Ziedonis D. A culturally adapted smoking cessation intervention for korean americans: A mediating effect of perceived family norm toward quitting. Journal of Immigrant and Minority Health. 2014:1–10. [PubMed: 23054547]
66. Kim SS, Fang H, DiFranza J, Ziedonis DM, Ma GX. Gender differences in the fagerström test for nicotine dependence in korean americans. Journal of smoking cessation. 2012; 7(01):31–36.
67. Houston TK, Sadasivam RS, Ford DE, Richman J, Ray MN, Allison JJ. The quit-primo provider-patient internet-delivered smoking cessation referral intervention: a cluster-randomized comparative effectiveness trial: study protocol. Implement Sci. 2010; 5:87. [PubMed: 21080972]
68. Zhang Z, Fang H, Wang H. Multiple imputation based clustering validation (miv) for big longitudinal trial data with missing values in ehealth. Journal of Medical System. 2016 accepted.
69. Little, RJ., Rubin, DB. Statistical analysis with missing data. John Wiley & Sons; 2014.
70. Andridge RR, Little RJ. A review of hot deck imputation for survey non-response. International Statistical Review. 2010; 78(1):40–64. [PubMed: 21743766]
71. Rubin DB. A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the sir algorithm. Journal of the American Statistical Association. 1987; 82(398):543–546.
72. Schafer JL. Multiple imputation: a primer. Statistical methods in medical research. 1999; 8(1):3–15. [PubMed: 10347857]
73. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological methods. 2002; 7(2):147. [PubMed: 12090408]

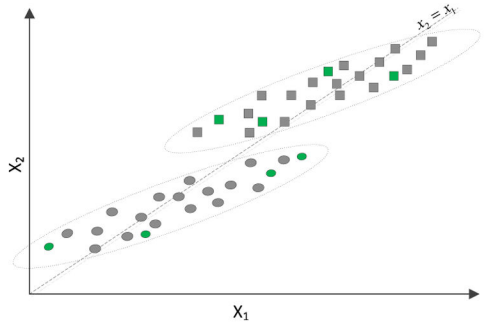


(a) A 2-dimensional 2-cluster dataset with simulated missingness

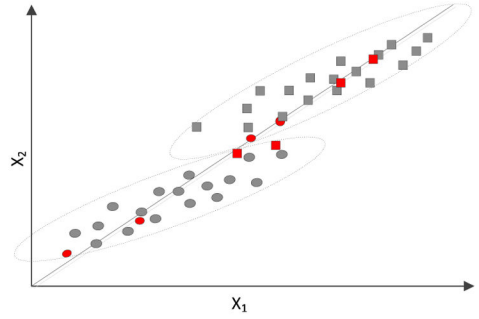


(b) Mean-imputation Results

Fig. 1.
An illustration of mean-imputation



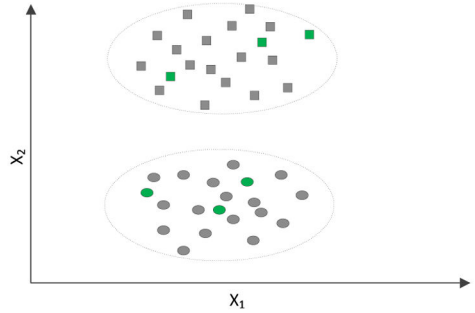
(a) The true two-dimensional incomplete data with two clusters



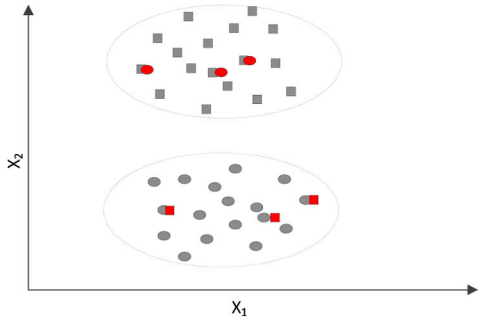
(b) Mistakenly clustered data after regression SI

Fig. 2.

Clustering accuracy of single regression imputation

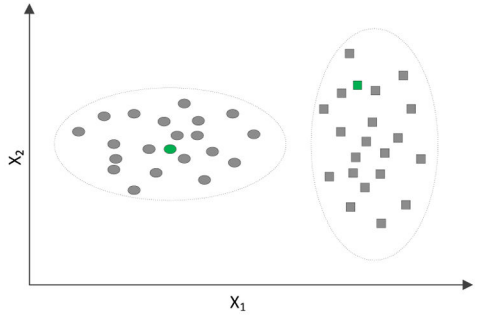


(a) The true two-dimensional incomplete data with two clusters

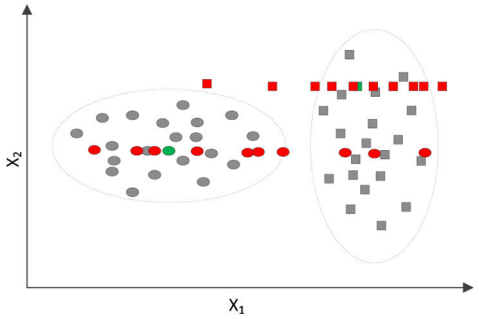


(b) Mistakenly clustered data after hotdeck SI clustering

Fig. 3.
Clustering accuracy of single hotdeck imputation



(a) The true two-dimensional incomplete data with two clusters



(b) Imputed data for the missing values

Fig. 4.
Clustering accuracy of multiple imputation

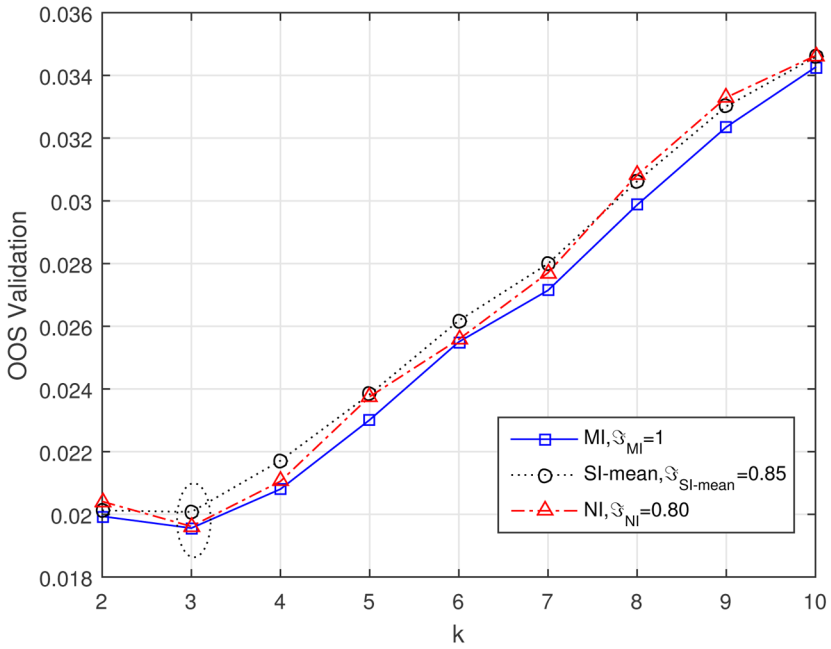


Fig. 5.
Performance of MI-, SI- and NI-clustering on TDTA data

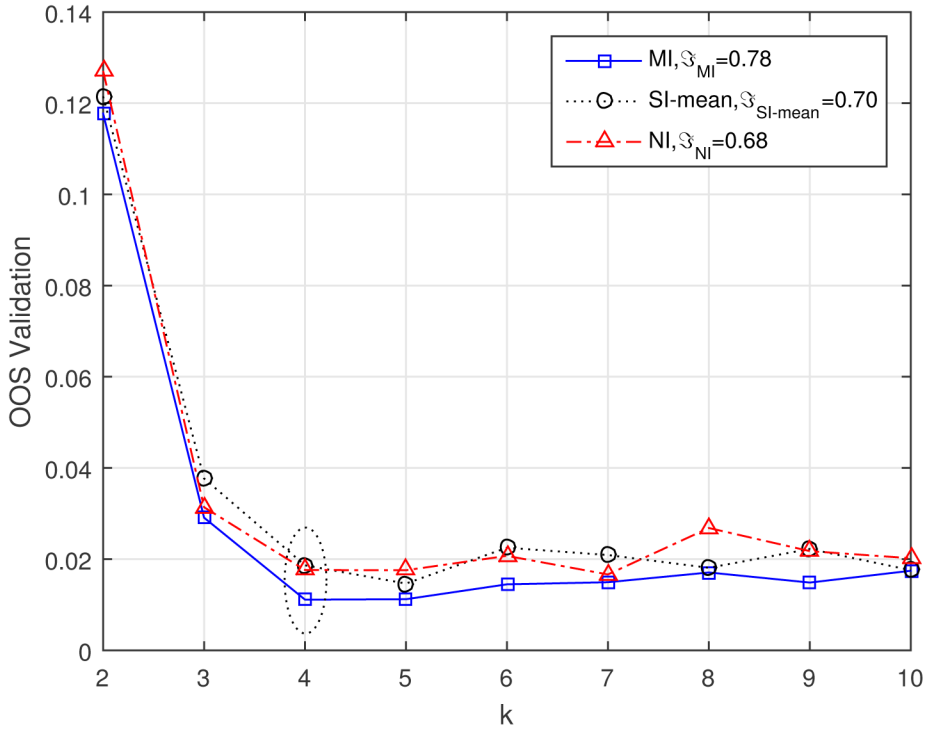
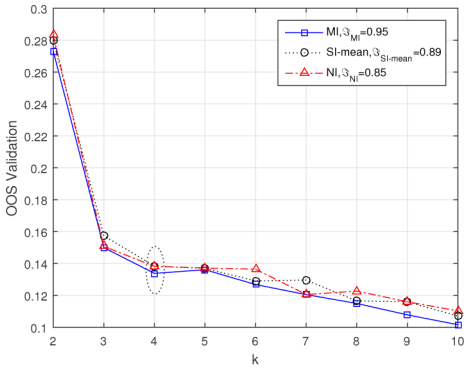
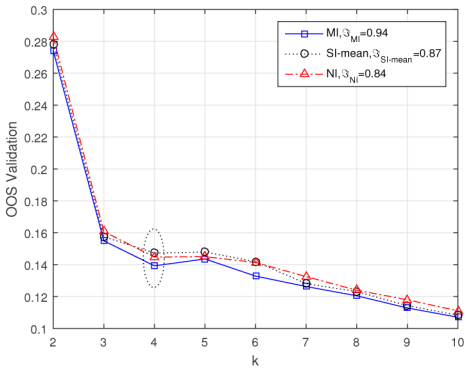


Fig. 6.
Performance of MI-, SI- and NI-clustering on QP data

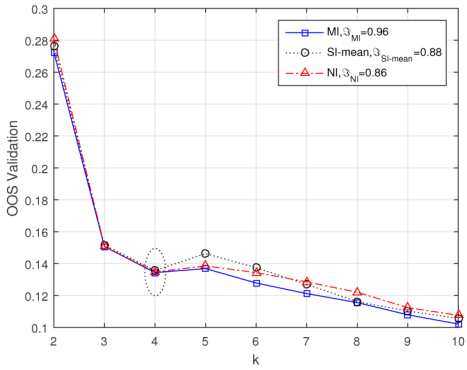


(a) Simulated MCAR data with 10% missingness

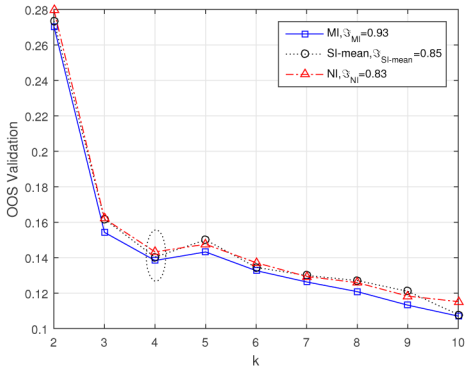


(b) Simulated MCAR data with 15% missingness

Fig. 7.
MI- vs. SI- and NI-clustering under MCAR

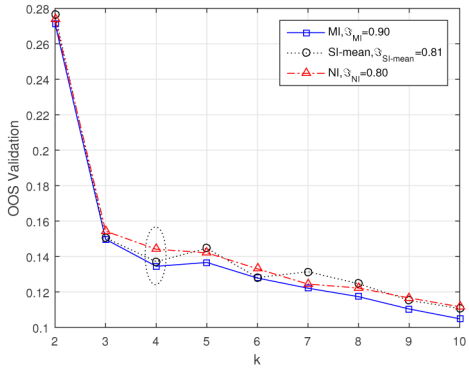


(a) Simulated MAR data with 10% missingness

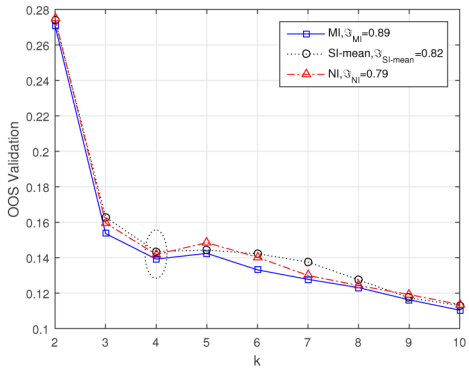


(b) Simulated MAR data with 15% missingness

Fig. 8.
MI- vs. SI- and NI-clustering under MAR



(a) Simulated MNAR data with 10% missingness



(b) Simulated MNAR data with 15% missingness

Fig. 9.
MI- vs. SI- and NI-clustering under MNAR

Table I

Notations

Symbols	Notations
X_{obs}, X_{mis}	Observed data and missing values
X	Dataset consists of X_{obs} and X_{mis}
M	Number of imputations
N	Number of cases in data X
N'	Number of cases in data X with missingness
$n_{control}$	Number of mislabeled control group members
$n_{intervention}$	Number of mislabeled intervention group members
$\mathfrak{J}$	Clustering accuracy
$\mathfrak{J}_{NI}$	Non-imputation clustering accuracy
$\mathfrak{J}_{SI\text{-}mean}$	Mean-imputation clustering accuracy
$\mathfrak{J}_{SI\text{-}regression}$	Regression-imputation clustering accuracy
$\mathfrak{J}_{SI\text{-}hotdeck}$	Hot deck-imputation clustering accuracy
$\mathfrak{J}_{MI}$	Multiple-imputation clustering accuracy
P, P'	Observed and unobserved attributes of X
CT	Termination clustering number
U	Fuzzy degree of cluster membership
V	Cluster centroids
n'	Number of patients with observed first attribute
r	Percentage of missing values
μ	Mean value
d	dimension of data X
k	Number of clusters

Table II

RE vs. r and M

M	r=10%	r=20%	r=30%	r=50%
3	0.968	0.938	0.910	0.857
5	0.980	0.962	0.934	0.910
10	0.990	0.980	0.971	0.952
20	0.995	0.990	0.985	0.976