PhUSE EU Connect 2018

Paper ML07

Handling Missing Data in Clinical Trials Using Topological Data Analysis

Sergey Glushakov, Intego Group, Maitland, FL, USA
Iryna Kotenko, Intego Group / Experis Clinical, Kharkiv, UKRAINE
Andrey Rekalo, Intego Group, Kharkiv, UKRAINE

ABSTRACT

Missing data is a major problem in clinical research. It can cause a loss of statistical power and often leads to biased estimates. The paper introduces a new methodology – topology-based clinical data mining (TCDM) – that can be used to deal with various issues related to missing data. TCDM is an integrated approach that combines biostatistics, topological data analysis, machine learning algorithms and data visualization tools. Its goal is to discover hidden patterns within a set of interrelated clinical outcomes by extracting comprehensive "topological maps" from the data. TCDM zooms in on robust patterns in the data that are not affected by random noise, and therefore topological data maps do not experience significant transformation if some data are missing at random. We show that TCDM can be applied effectively to several problems, including automated selection of a suitable imputation method, validation of imputation methods.

INTRODUCTION

Clinical trials datasets invariably contain missing values. The problem then is to recover the most probable values through an appropriate data imputation process before proceeding to the analysis of the dataset. Over the years, many data imputation methods have been developed, but the tools for effectively selecting and validating a suitable method for a given dataset are not readily available to practitioners.

This paper introduces a new methodology – Topology-based Clinical Data Mining (TCDM) – that can be used to deal with various issues related to missing data. By suggesting a principled algorithmic solution to the data imputation problem, TCDM may play an important role in data pre-processing and, hence, in the selection of a suitable statistical model for the analysis of clinical dataset.

TCDM is an integrated approach that combines topological data analysis, biostatistics, machine learning algorithms, and data visualization tools. It allows the extraction of comprehensive topological data maps which provide compressed graphical representations of a multidimensional set of interrelated clinical outcomes. TCDM focuses on topological properties of data, i.e. robust data patterns and relationships that remain invariant under relatively small perturbations of the data.

A topological data map representing a clinical dataset is a graph in which every node corresponds to a single patient, while patients who are sufficiently similar (in terms of predefined outcomes) are connected with an edge. The TCDM focus on topological properties helps to identify the underlying geometry of a clinical dataset and to determine whether imputed values filled in with a particular imputation method materially distort the geometry of the original data.

A group of related mathematical methods that are collectively known as Topological Data Analysis (TDA) has recently been applied in different branches of bioinformatics, epidemiology, neuroscience, and oncology with promising results. TDA is a rapidly expanding field that is being actively developed by research teams in leading academic centers both in the US and Europe, including renowned academic centers such as Stanford, Duke, UPenn, Princeton Neuroscience Institute, INRIA (France), among many others. TDA methods are based on the underlying idea of using topology – the mathematical study of qualitative properties of space and spatial relations – to detect and display hidden robust relationships in complex datasets. Thus far, this idea has been successfully applied to discover a coherent subgroup of breast cancer patients with 100% survival, which is characterized by a unique molecular signature [1]; to reveal unexpected statistically significant patterns in traumatic brain injury and spinal cord injuries [2]; to distinguish resilience to malaria in human populations [3]; and to identify in silico drug leads from a diverse library of compounds [4].

The paper is organized as follows. Section 2 contains a brief overview of missing data mechanisms and standard approaches to deal with missing values. It proceeds to describe how methods of Topological Data Analysis can be adopted to evaluate and validate imputation methods based on the geometry of a dataset. Section 3 outlines a case study where the topology-based methods are applied to the problem of automated selection of a suitable imputation method for a Treatment Effectiveness Assessment questionnaire within a real clinical trial. Section 4 concludes.

2. METHODS

2.1. TYPES OF MISSING DATA

According to Rubin [5], missing data mechanisms may be categorized into three groups based on the assumption that, for each missing value, one can associate with it a conditional probability of the value being missing depending on the data.

The first missing data mechanism relates to data missing completely at random (MCAR) when the probability of being missing for a data record does not depend on any data, either available or missing. In case of MCAR, reasons that lead to missing of data are irrelevant to the data. Therefore, complexities resulting from the missingness, except for the obvious loss of information, may be ignored. A random sample of a population, in which each subject has the same likelihood of being included in the sample, may illustrate MCAR. Specifically, the data related to subjects that are not present in the sample are MCAR. Although the MCAR assumption is convenient for practical purposes, it may be unrealistic for a given set of data.

The second mechanism is concerned with data missing at random (MAR) when the probability of being missing for a data record may depend on the available (non-missing) data. MAR may be illustrated by a sample taken from a population, in which the probability of inclusion into the sample depends on some known criteria. As compared to MCAR, the assumption of MAR is more general and realistic. The MAR assumption is usually considered first in missing data analysis.

The third missing data mechanism deals with the missing not at random (MNAR) scenario when the probability of being missing for a data record may depend on both available and missing data. For example, MNAR happens when those participants of a clinical trial who have more serious condition provide a response less often than the rest. MNAR can be handled by determining or modeling the reasons explaining why the data are missing and analyzing the sensitivity of the results to various conditions.

2.2. THE INDUSTRY OUTLOOK ON THE MISSING DATA ANALYSIS

Traditional methods for handling missing data in longitudinal clinical trials include complete-case analysis, last observation carried forward (LOCF), baseline observation carried forward (BOCF), and some other simple forms of data imputation. The conclusions drawn based on the results of the analysis are usually made without considering the implications of assumptions made in the course of the analysis, even though a plenty of studies addressing the related issues are currently available.

A complete-case analysis is a simple method that involves consideration of subjects (cases) that comprise a complete set of values. Subjects with missing values are eliminated from the analysis. Statistical software packages, such as SAS and SPSS, apply the complete-case analysis as a default method.

The complete-case analysis is a convenient method that may be advantageously used, for example, if the data are missing completely at random (MCAR). In this scenario, researchers may obtain correct estimates of standard errors and significance levels by considering only complete case data. However, in situations, when large portions of data are missing and there are many variables in a dataset, the complete-case approach may result in a very small sample of subjects which are available for analysis. Additionally, the estimates of means, correlations, and regression coefficients may be substantially biased when the data are not MCAR [6].

The LOCF method imputes each missing value with the last observed value from the same subject. To apply LOCF, it is assumed that missing values are MCAR and the last measured measurement of a given subject remains the same for the rest of the clinical trial. This method has been shown to have numerous drawbacks (see, e.g. [7-9]). However, LOCF is still one of the most frequently used data imputation method in statistical studies (e.g., a search in Google Scholar with “LOCF” as a key word shows approximately 1470 results for 2017). Thus, though simple imputation methods are based on unrealistic assumptions and have questionable statistical validity, these methods remain widely used in analysis of clinical trials data.

The Panel on Handling Missing Data in Clinical Trials, under the authority of the FDA, issued a report [10] which recommended using advanced imputation methods such as maximum likelihood, multiple imputation, Bayesian methods, and methods based on generalized estimating equations. The Panel also recommended that single imputation methods, such as LOCF and BOCF, should not be used as the primary approach unless the assumptions that underlie this decision are scientifically justified.

Unfortunately, the advanced imputation methods are often complicated and time- consuming, and the complete-case analysis, LOCF and BOCF methods remain the most commonly used by practitioners despite their well-documented deficiencies. Meanwhile, the Panel emphasized that progress was needed in several important areas, including the development of robust methodologies and analytics tools that can support coherent missing data analyses.

TCDM can tackle missing data problems by effectively visualizing how a given data imputation method conforms to the geometry of the complete case data subset. This geometric approach also allows one to select a suitable data imputation method in an automatic, objective way and, subsequently, to fine-tune the parameters for that method.

2.3. ROBUST GEOMETRIC PROPERTIES OF DATASETS

For the illustrative purposes, a simple two dimensional dataset was constructed where data points were arranged in a “zero-like” shape. TCDM was applied to this dataset to build a topological data map in the form of a graph in which every node corresponds to a single data point.

In order to show robustness of the topological approach, some data points from the dataset were intentionally omitted at random, and additional graphs were built for the modified datasets where 50% and 90% of the original data points were missing (see Figure 1).

The graphs show certain geometrical stability even in the case of 90% missingness. The shape of graphs built on the remaining data points is structurally similar to the shape of the graph corresponding to the complete dataset. Therefore, in this example, topological data maps representing a relatively small portion of the data still have similar shape to the graph representing a complete dataset.

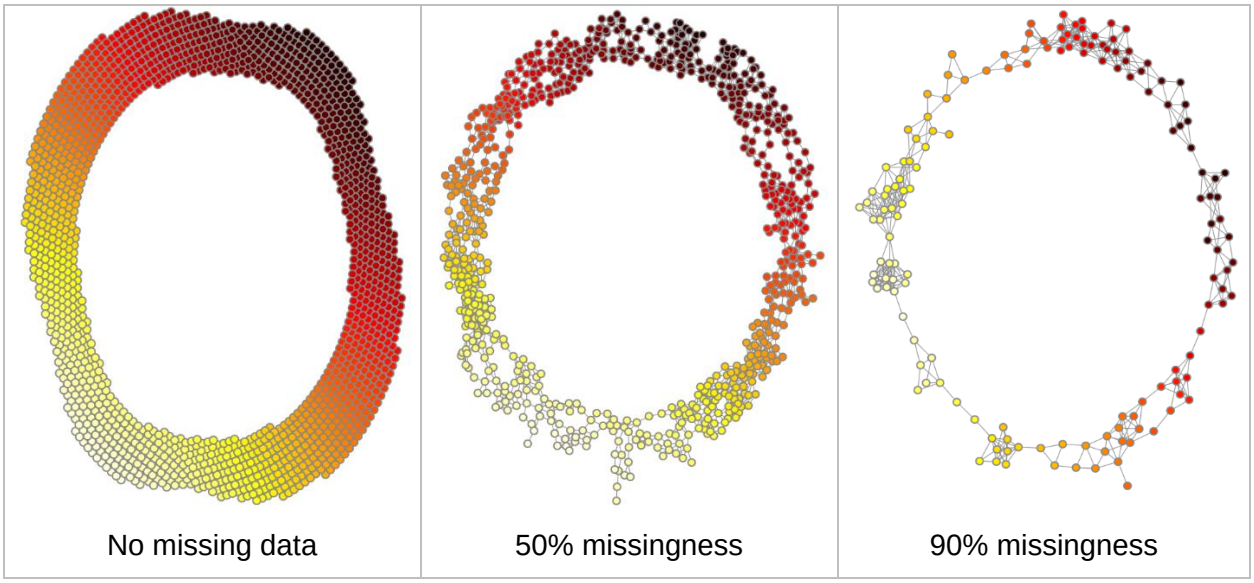


Figure 1. Topological data maps representing a dataset with varied proportion of randomly missing data

Graphs produced by the TCDM algorithm for a complete dataset (left panel) and datasets where 50% and 90% of data points are missing (middle and right panels, respectively). This example illustrates that even with 90% of the data missing, the cyclic shape of the dataset is preserved by the corresponding topological data map

2.4. WHAT IS A TOPOLOGICAL DATA MAP?

The core idea of TCDM relies on a visual discovery of subgroups of related patents in a topological data map (see Figure 2) that retains the relevant information about a dataset in a compact and efficient manner. To be considered for further analysis, a topological data map should meet certain requirements:

- **Each node represents patient** – a topological data map is a graphical representation of the dataset in which each node represents an individual.
- **Similar nodes are connected** – two nodes representing similar patients (in terms of a predefined set of clinical outcomes) are connected with an edge.
- **Coloring focused on specific outcomes** – the color of the nodes helps to highlight emerging patterns in data and to identify subgroups of patients related to the distribution of a variable of interest.
- **Visual discovery of subgroups** – clusters or "communities" of nodes on a topological data map reflect a segmentation of patients that may indicate robust patterns within the data.

To construct a visual representation of clinical trial data, the dataset in CDISC format is pre-processed using proprietary algorithms that have been developed to deal with data-specific issues, such as proper scaling of numerical variables, conversion of categorical variables, and others. At the initial stage, a primary dataset needs to be determined where each row in a data table corresponds to a unique patient or volunteer who participated in the clinical study while the columns represent either observational variables (outcomes) such as safety and efficacy biomarkers or predictors such as demographic attributes, medical history, interventions, etc. The resulting dataset is further processed by a computational platform to construct a visualization of the observational variables represented by a topological data map.

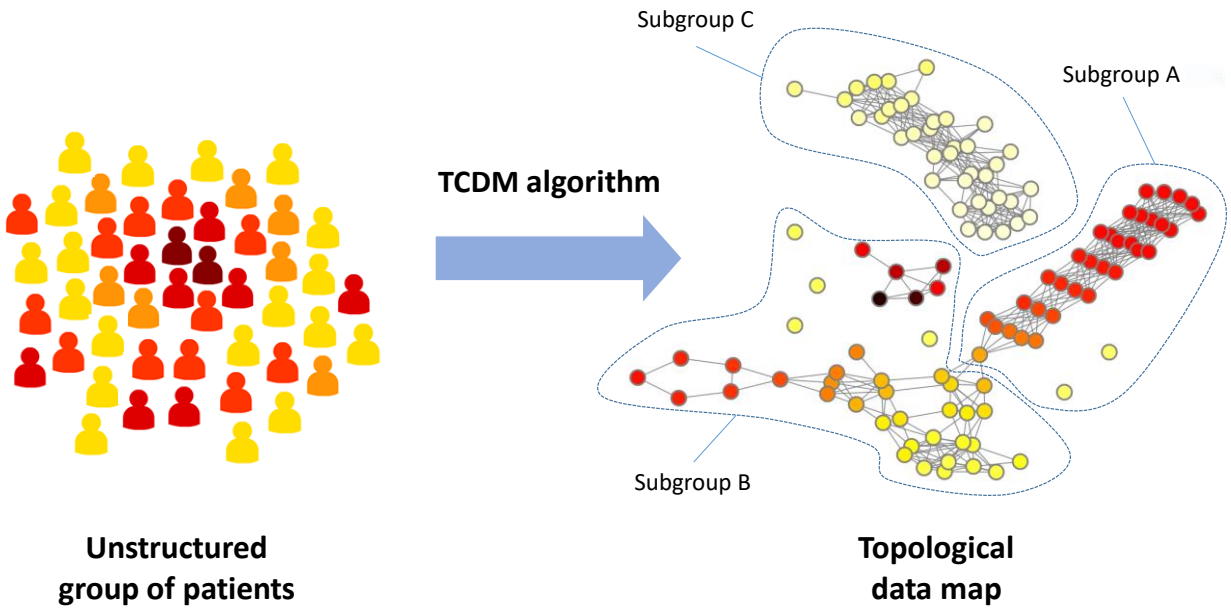


Figure 2. Discovery of multivariate patterns in clinical trial outcomes

The topological data map represents groups of patients structured according to similarity of clinical outcomes

TCDM can deal with a variety of numerical and categorical outcomes:

- **Interrelated biomarkers** – e.g. patients' vital signs or basic metabolic panel results on a specific day of study.
- **Series of repeated measurements** – e.g. weekly hemoglobin levels during chemotherapy in oncological patients.
- **Questionnaire data** – binary, nominal or ordinal responses to the items of a questionnaire, aggregate scores, etc.

After a topological map is constructed based on selected outcomes, the researcher then visually explores the data map with the purpose of discovering interesting subgroups within the data. For example, isolated components of a data map or highly interlinked groups of nodes that form communities may indicate meaningful relationships within the dataset.

2.5. WORKFLOW FOR GEOMETRIC EVALUATION OF IMPUTATION METHODS

The core concept of TCDM relies on the extraction of robust topological maps of a dataset that would not be significantly distorted if some data were missing at random. As such, the first step of the workflow involves statistical tests allowing to identify if data is missing at random (see Figure 3).

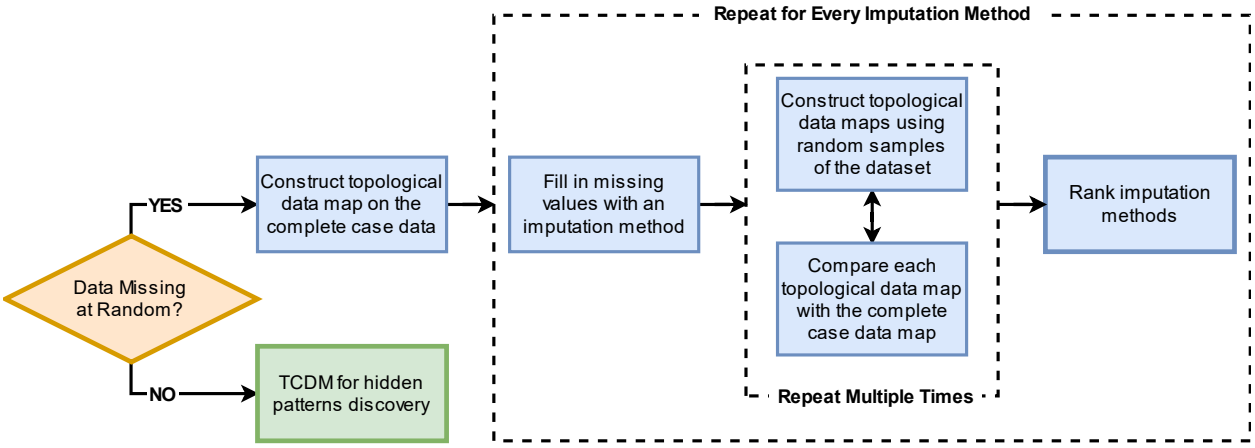


Figure 3. TCDM workflow for evaluation of standard imputation methods

TCDM can effectively help to validate different data imputation methods by comparing topological data maps built for complete case data and a resampled dataset with imputed values

If the data is found to be missing not at random, TCDM can be used to discover hidden relationships that might explain the missingness patterns of the data and provide additional insights into the clinical dataset. The patterns of data missingness can be integrated into the dataset as additional outcomes in order to construct an extended topological data map. After that, researchers can combine a visual exploration of the data map together with statistical analysis of predictors to identify possible reasons for the observed patterns of missingness. This approach of using TCDM to discover hidden relationships within a dataset was described in our PhUSE 2017 paper [11].

If the data are missing at random, a variety of standard methods may be applied to impute the missing values. After the imputation step is completed, a topological data map for the dataset with imputed values can be constructed. Then, researchers can determine whether the geometric structure of this data map is similar to the structure of the data map extracted from the complete case dataset (i.e. the dataset obtained from the original dataset by listwise deletion of rows containing missing values). If these two data maps have similar geometry, this would imply that the data imputed with the particular imputation method conforms to the geometry of the complete case dataset. Thus, the researchers may conclude that the imputation method works well for this specific dataset in terms of the underlying data geometry.

2.6. AUTOMATED SELECTION OF AN IMPUTATION METHOD DRIVEN BY THE DATA GEOMETRY

If the data within a clinical dataset is found to be missing at random, the TCDM computational platform can be used first to construct a topological data map for the complete case dataset (see Figure 3). For the purpose of analysis, the original dataset is represented as a table in which every row contains the data for a single patient. Therefore, some rows are complete, while some may have missing values. The complete case dataset comprises only the rows with no missing values, and its corresponding topological data map is considered the *reference graph* of the dataset.

Then, researchers can apply one of the standard methods of data imputation to fill in the missing values. To evaluate a given imputation method, a substantial number of subsamples is selected from the dataset with imputed values. Each subsample is selected at random and has the same size (i.e. same numbers of rows and columns) as the complete case dataset (see Figure 4). This means that subsampled datasets could include both original rows without missing values and rows with imputed values.

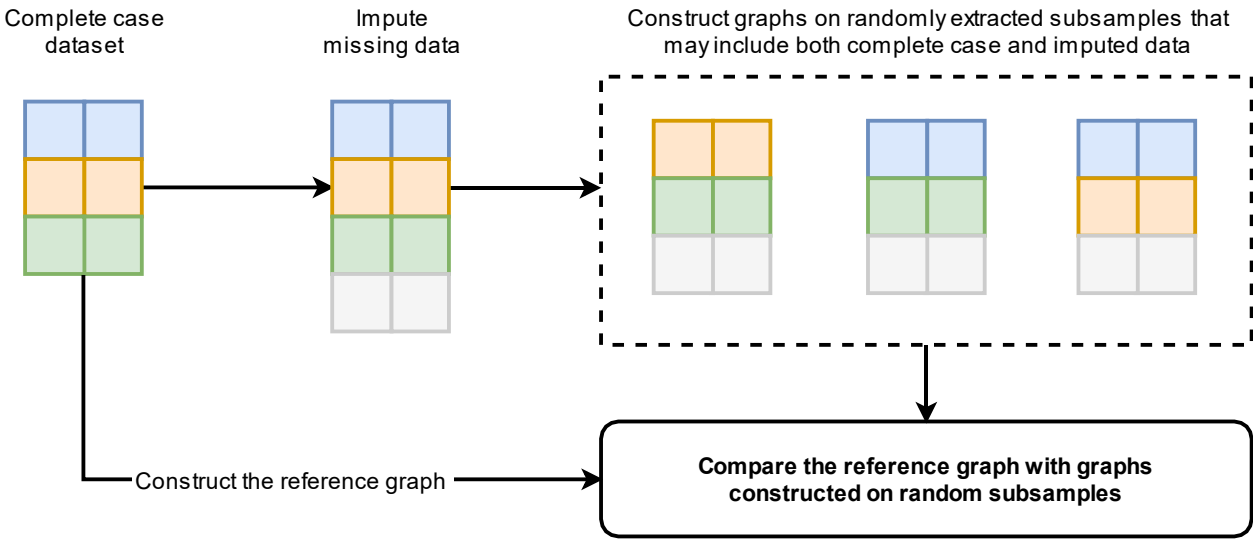


Figure 4. Generation of subsampled datasets with imputed values

Graphs extracted from the dataset with imputed values are compared with a reference graph extracted from the complete case data

For each subsample, a topological data map is constructed and compared with the reference graph to determine the degree of similarity. To do this, researchers calculate the pairwise distances between graphs corresponding to subsamples and the reference graph. The smaller the distance to the reference graph, the more similarity there is between the two graphs, and hence, between the complete case and subsampled datasets. Researchers can repeat this process for different imputation methods and compare the respective average distances in order to rank the imputation methods and ultimately select the most appropriate one for the original dataset under consideration.

3. RESULTS

TCDM was applied to the evaluation of several popular data imputation methods for the Treatment Effectiveness Assessment (Health) (TEAHL) dataset from a clinical study that was conducted with the support of the NIDA Clinical Trials Network (the study data are available at https://datashare.nida.nih.gov/study/nida-ctn-0048). The TEAHL questionnaire asks a patient to express the extent of changes for the better from his/her involvement in the clinical study. For each area, the patient is asked to think about how things have become better and to circle the results on a 10-point scale, where 1 represents “not better at all” and 10 represents “very much better.” In this particular clinical study, the TEAHL dataset comprised data from 302 patients with 5 observations for each patient; 248 patients had no missing values (complete case data), and 104, or about 7 %, out of 1510 values in total were missing. The baseline observations were available for all patients.

The analysis commenced with the extraction of a reference graph for the complete case dataset comprising 248 patients. Then, missing values were imputed using several standard data imputation methods, namely, LOCF , k-nearest neighbors (*k*-NN), and Multivariate Imputation by Chained Equations (MICE, [12]). For each imputation method, 1000 subsamples consisting of 248 patients each were randomly selected from the original dataset. All subsamples included both patients with and without missing values that were subsequently imputed using one of the methods.

Each topological data map constructed for a subsample was compared with the reference graph to determine geometric similarity, or distance, between the graphs (see Figure 4), as described in the previous section. The effectiveness of LOCF, *k*-NN, and MICE imputation methods was assessed according to the average distance between the graphs corresponding to subsampled datasets and the reference graph.

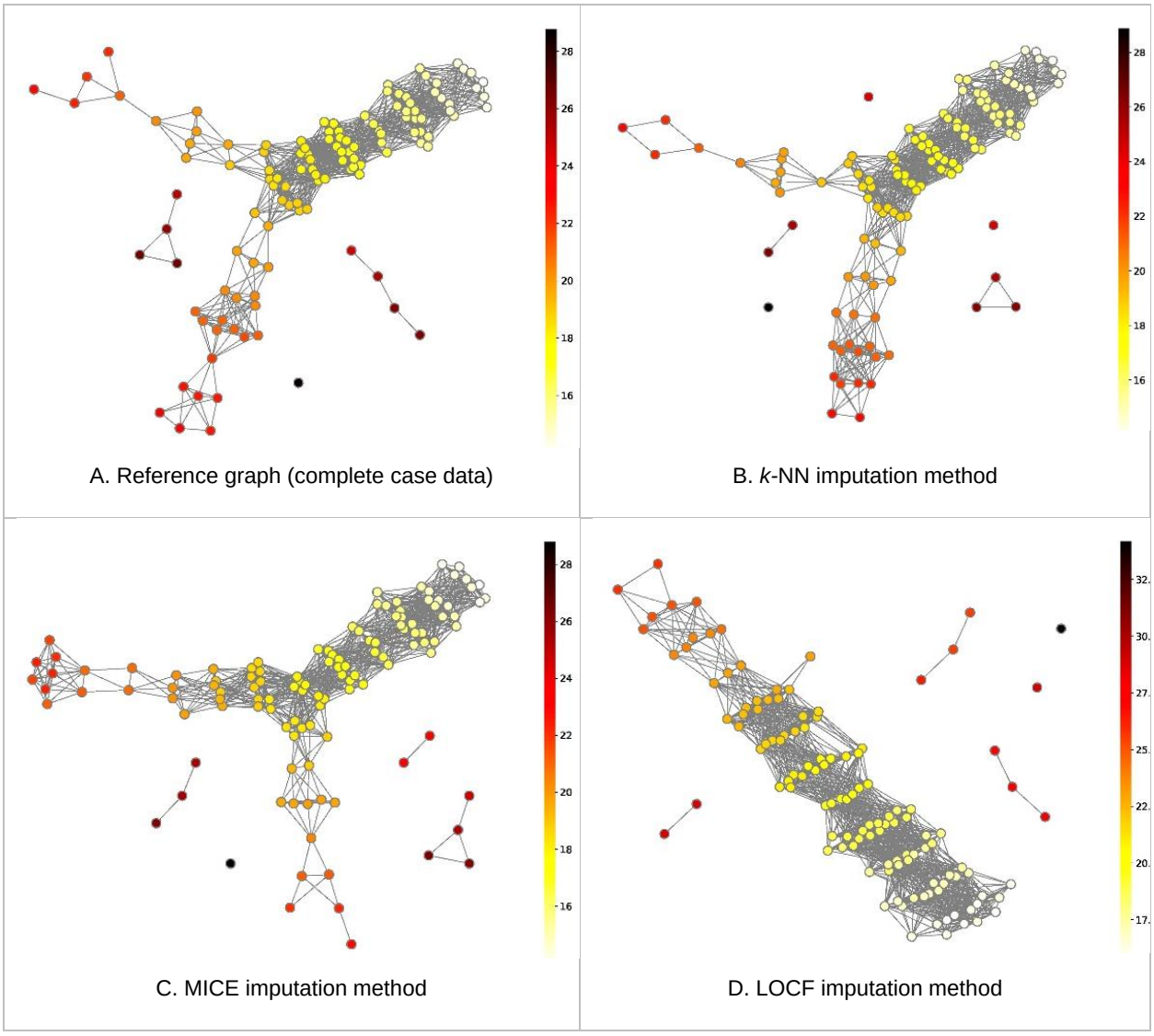
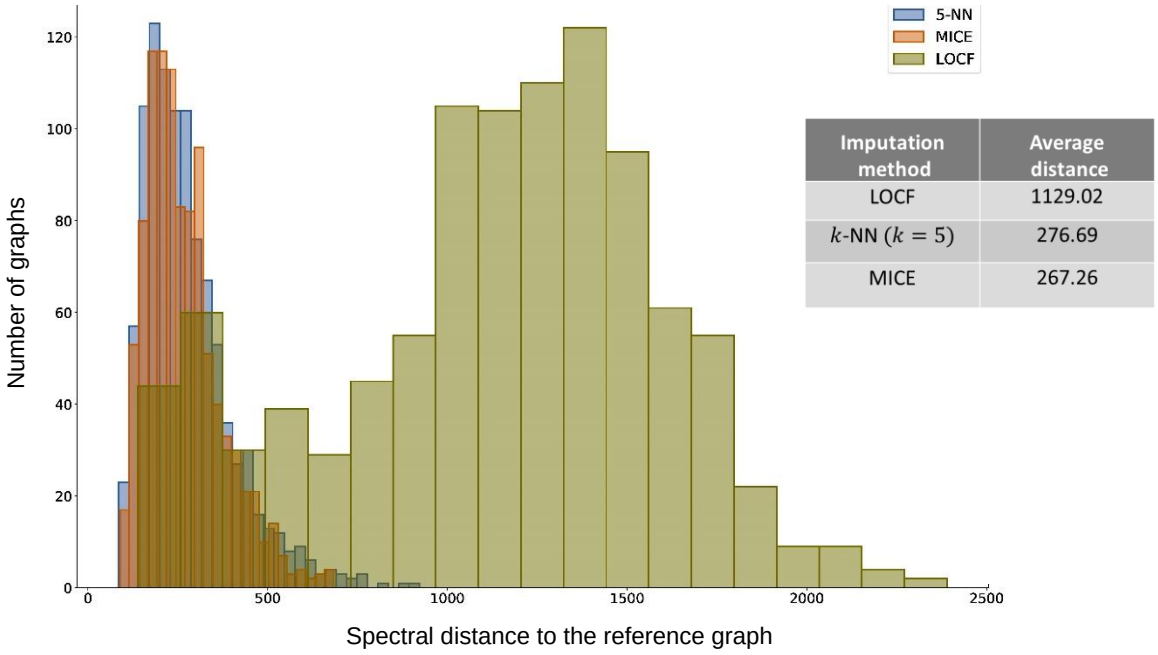


Figure 4. Topological data maps corresponding to the complete case and subsampled datasets where missing values were imputed with 3 different methods

This figure illustrates graphs constructed for a TEAHL complete case dataset (upper left panel) and subsampled datasets with missing values that were subsequently imputed using *k*-NN, MICE, and LOCF imputation methods accordingly (remaining panels). For illustrative purposes, the number of patients within each dataset was reduced to 120 (from the original 248)

In addition, researchers varied the parameters of the *k*-NN imputation algorithm to optimize its performance. This analysis revealed that for the TEAHL dataset, the *k*-NN algorithm with $k = 5$ lead to data maps that were the closest on average to the reference graph.

In the final stage, researchers determined whether the differences between the average distances among the data imputation methods were statistically significant. *k*-NN and MICE imputations resulted in graphs that were the closest to the reference graph (the average distances were 276.69 and 267.26, respectively) However, the difference between the corresponding average distances was not statistically significant at 0.05 level (Welch's t-test, p-value was 0.068)., At the same time, both *k*-NN (k=5) and MICE proved to be much more suitable data imputation methods than LOCF for the TEAHL dataset in terms of preserving the complete case data geometry (see Figure 5).



Imputation method	Average distance
LOCF	1129.02
k-NN ($k = 5$)	276.69
MICE	267.26

Figure 5. Distribution of distances to the reference graph for 3 different imputation methods

Graphs for the data imputed by the *k*-NN (k=5) and MICE methods were significantly closer to the reference (complete case data) graph on average than were graphs for the data imputed by the LOCF method

4. CONCLUSION

Owing to its capability to visualize clinical datasets by extracting robust topological data maps, Topology-based Clinical Data Mining is a viable solution to the problem of principled evaluation of imputation methods. TCDM methodology is based on the geometry of the underlying data and does not rely on any statistical assumptions concerning the data distribution. By comparing representative topological data maps generated for the complete case data (the reference graph) and data with imputed values (graphs constructed on random subsamples of the dataset after imputation), researchers can objectively select, evaluate and calibrate a suitable imputation method for a specific dataset.

REFERENCES

[1] Nicolau, M., Levine, A.J., Carlsson, G. “Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival.” Proc Nat Acad Sci. 108 (2011): 7265-270.

[2] Nielson, J. L. et al. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. Nat. Commun. 6:8581 doi: 10.1038/ncomms9581 (2015).

[3] Torres, B.Y., et al. “Tracking Resilience to Infections by Mapping Disease Space.” PLOS Biology E1002494 14.4 (2016): n. pag. 10.1371/journal.pbio.1002494.

[4] Alagappan, M. et al. “A Multimodal Data Analysis Approach for Targeted Drug Discovery Involving Topological Data Analysis (TDA).” Advances in Experimental Medicine and Biology Tumor Microenvironment 899 (2016): 253-268. 10.1007/978-3-319-26666-4_15. Springer.

[5] Rubin, D.B. “Inference and Missing Data.” Biometrika 63 (1976): 581–590.

[6] Donner, A. “The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values.” Am Stat. 36 (1982): 378–381.

[7] Lachin, J.M. “Fallacies of last observation carried forward analyses.” Clin Trials 13 (2016):161-168.

[8] Kenward M.G., Molenberghs, G. “Last Observation Carried Forward: A Crystal Ball?” J Biopharm Stat, 19 (2009): 872–888.

[9] Molnar, F.J., Hutton, B., Fergusson, D. “Does analysis using “last observation carried forward” introduce bias in dementia research?” Can Med Assoc J 179 (2008): 751–753.

[10] The Prevention and Treatment of Missing Data in Clinical Trials, National Research Council (US) Panel on Handling Missing Data in Clinical Trials. Washington: National Academies Press; 2010.

[11] Glushakov, S., Kotenko, I., Rekalo, A.“Topology-based Clinical Data Mining: Identifying Hidden Patterns in Clinical Datasets” PhUSE 2017, paper TT10.

[12] van Buuren, S. “Multiple imputation of discrete and continuous data by fully conditional specification.” Stat Methods Med Res. 16 (3) (2007): 219-242.

ACKNOWLEDGMENTS

We would like to acknowledge Bogdan Chornomaz (Kharkiv National University, Ukraine / Vanderbilt University, United States) and Kostiantyn Drach (Kharkiv National University, Ukraine / Jacobs University Bremen, Germany) for being core members of the research team and for the significant contribution they made to the development of the mathematical foundation of the TCDM methodology. Without you this research would not have been possible.

We thank our colleague Victoria Shevtsova (Intego Group), who greatly assisted the research team to execute a variety of experiments using clinical trial data and made a significant contribution to the development of the software prototype the researchers employ to perform the data analysis.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the group of authors at:

Contact: Sergey Glushakov
Company: Intego Group
Address: 555 Winderley Place, Ste. 129, Maitland, FL 32751
Work Phone: 407.641.4730
Email: sergey.glushakov@intego-group.com
Web: www.intego-group.com