PharmaSUG 2024 - Paper DV-313

Visual Discovery in Risk-Based Monitoring Using Topological Models

Kostiantyn Drach, University of Barcelona/Intego Group, Barcelona, Spain
Iryna Kotenko, Intego Group, Kharkiv, Ukraine

ABSTRACT

A strong endorsement from the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) for efficient oversight of clinical investigations leads to a crucial role of risk-based monitoring (RBM) in new drug development. The main goal of the RBM approach is to offer strategic and effective ways to allocate resources across a study based on several key indicators, such as data criticality, patient safety, protocol compliance, and others. One of the components in RBM is the Clinical Trial Site Monitoring which presents a significant challenge as it requires timely insights generation from the data coming from sites in almost real-time. As monitoring sites' activities is an important task to ensure protocol compliance and safety of patients and the resulting drugs, in this paper, we introduce how to aim it with a visualization approach using topological models of the data. We represent the clinical data using a graph (a topological model) that captures the geometric properties of complex data and where each node corresponds to a clinical trial subject, while two nodes are connected with an edge if these two subjects have similar outcomes/indicators of interest. A variety of graphs can be generated depending on indicators of interest. Those with robust geometric structures may further be analyzed using interactive operations and various machine learning (ML) algorithms. Using the topological models, the researcher can easily highlight data coming from specific sites and further analyze problematic pieces of data. In an experiment, we demonstrate this visual discovery approach compared to standard statistical methods.

1. INTRODUCTION

Recently, the growth and complexity of clinical trials have presented challenges in oversight due to increased variability in investigator experience, site infrastructure, treatment choices, and healthcare standards. Nevertheless, advancements in computer systems, electronic records, statistical assessments as well as Artificial Intelligence (AI)/ML usage offer opportunities for alternative monitoring approaches, such as centralized monitoring, to enhance the quality and efficiency of sponsor oversight. According to [1], RBM is the process of ensuring the quality of clinical trials by identifying, assessing, monitoring, and mitigating the risks that could affect the quality or safety of a study. The FDA encourages sponsors to develop monitoring plans that address challenges emphasizing a risk-based approach focused on preventing or mitigating significant risks to data quality and processes critical to human subject protection and trial integrity.

The key components of RBM, as outlined in [2], are:

- key risk indicators (KRIs);
- centralized monitoring;
- of-site/remote-site monitoring;
- reduced source data verification (SDV);
- reduced source document review (SDR).

One of the key components of RBM is its use of centralized monitoring techniques. As opposed to on-site monitoring based on 100% SDV, centralized monitoring offers a number of advantages:

- **fewer errors** due to less manual work;
- **lower cost due to** reducing the frequency and extent of on-site monitoring and auditing only the sites where problems are most likely to occur;

- **better analysis and cross-site comparison,** since centralized monitoring also allows us to compare data between sites and to identify potentially fraudulent, inaccurate, or biased data.

The FDA has provided some detailed guidance on how to prepare a monitoring plan [1], although execution of the plan for specific clinical trials is still challenging and requires the search for new approaches. We propose some visual solutions in RBM based on *topological data analysis* (TDA) approach in combination with other AI/ML techniques (see Sections 2 and 3 for methods description). These solutions can help to supplement the statistical by-site data processing, identify some KRIs, as well as can be used to reveal the problematic sites and provide alerts to perform targeted on-site investigation. These solutions are flexible enough to account for any equity and diversity within the study.

To demonstrate the working prototypes of solutions, we use a publicly available National Institute on Drug Abuse (NIDA) dataset [3]. The original NIDA experiment investigated if the buprenorphine/naloxone combination tablet can be effectively used to treat patients with opiate dependence. A total of 582 participants on 38 centers were recruited. Study participation lasted 9 to 12 weeks for patients who successfully achieved detoxification and up to 12 months for patients requiring longer buprenorphine treatment. On baseline and during the trial, various data concerning general health, vital signs, drug addiction, family history, psychological health, employment status, and other information (total over 280 variables) were gathered, as well as treatment results were indicated. The variety of data allowed us to carry out the experiments, which are described in Sections 3 and 4.

2. TOPOLOGICAL DATA ANALYSIS

TDA is a relatively novel approach of building visual representations of complex datasets. This analysis yields the extraction of comprehensive graphs from a dataset to provide a compressed graphical representation of a multidimensional set of interrelated outcomes. When applied to clinical data, these graphs consist of nodes corresponding to patients participating in a study and edges connecting those patients who share similarities in terms of study outcomes or other indicators of interest. This section focuses on the concept of the geometric properties of a dataset to understand how graphs can be extracted from complex datasets and further modern ML algorithms and visual exploration techniques can be used to detect subgroups of patients that share similarities.

2.1. TOPOLOGY AND DATA MINING

Topology is a field of mathematics that deals with the properties of objects that remain invariant under continuous deformation. Imagine a surface that is made of a very thin and elastic material. The surface can be bent, stretched, or crumpled in any way; however, it cannot be torn and its parts cannot be glued together. As the surface is deformed, it changes in many ways, but some properties remain the same. The idea underpinning topology is that some geometric properties depend not on the exact shape of an object but, rather, on how parts of the object are combined.

As a simple example, consider geometric figures on the plane representing the numerical digits 0, 1, 2, … 9. For a topologist, various representations of the digit 0 are equivalent since they can all be continuously transformed into each other without cutting or gluing (see **Figure 1 a-d**). It is possible to change the size, thickness, or slope of the digit 0 through continuous deformation; however, one property remains invariant: the object separates the plane into two regions, namely an interior region and an exterior region. At the same time, 0 is not topologically equivalent to 1 or 8: 1 does not encircle a region and 8 contains two holes (see **Figure 1 e**). The topological classification of the digits 0, 1, 2, … 9 results in the following five classes:

{0}, {1, 2, 3, 5, 7}, {4}, {6, 9}, {8}.

The digits in any of the classes are topologically identical, but no two digits taken from distinct classes are equivalent from the topological point of view.

a) b) c) d) e)

Figure 1. Different representations of the digit 0 (a-d) are topologically equivalent. All share a common topological property: they divide the plane into an interior region and an exterior region. The digit e) is not equivalent to 0 since it encloses two interior regions

The number of holes in a geometric object is a basic topological property. Another significant property is connectedness. Intuitively, an object is connected if it consists of a single piece. For example, the curve representing 0 is connected; if any two points are removed from it, it becomes disconnected. Pieces of a disconnected object that are themselves connected are referred to as connected components. In the mathematical study of topology, all of these intuitive concepts are examined on a rigorous basis and generalized to higher dimensions.

Topology deals with abstract mathematical entities, such as curves and surfaces, that consist of an infinite number of points. In practice, however, all datasets are necessarily finite. Recently, a new field has emerged at the crossroads of topology and data science. TDA aims to extract topological data, i.e., qualitative information, from finite sets of data points. It involves exploring datasets (viewed as finite clouds of points in multidimensional space) at multiple scales or resolutions, from fine- to coarse-grained. Given a complex dataset, TDA can be used to extrapolate the underlying topology and build a compressed yet comprehensive topological summary of the dataset. TDA exploits various methods and algorithms stemming from computational topology and geometry, statistics, and data mining. For detailed expositions of the mathematical theories that underpin TDA and certain applications in biology, see [4], [5], [6], and the references therein.

Topology was originally developed to distinguish between the qualitative properties of geometric objects. It can be used in conjunction with the usual data-analytic tools for the following tasks:

1. **Characterization and classification**. Topological features succinctly express qualitative characteristics. In particular, the number of connected components of an object is of importance for classification.
2. **Integration and simplification**. Topology is focused on global properties. From the topological perspective, a straight line and a circle are locally indistinguishable; however, they are not equivalent if they are considered as a whole. Topology offers a toolbox to integrate local information about an object into a global summary. Thus, topology can provide the researcher with a natural "big-picture" view of complex, multidimensional data.
3. **Features extraction**. Topological properties are stable. The number of components or holes is likely to persist under small perturbations or measurement errors. This is essential in data mining applications because real data are always noisy.

2.2. TOPOLOGICAL DATA MODEL

In the context of clinical research, a dataset under study is typically a table of variables in a particular clinical trial. The table rows correspond to individual participants in the clinical trial, and the columns contain information on specific variable measures of interest, such as lab tests, vitals, questionnaires, etc.

Each row of this table can be depicted as a vector of variables describing a particular patient; this vector can also be visualized as a point in a multidimensional space. However, visualizing the cloud of points

representing all participants in a clinical study becomes challenging when the dimensionality exceeds three.

To facilitate the visualization of data in multidimensional scenarios, dimensionality reduction methods prove to be particularly valuable. These methods reduce the number of variables (represented by columns of data) needed to describe each data point while preserving the underlying data structure.

The dimensionality reduction methods simplify the visualization of datasets with numerous columns. The commonly used dimensionality reduction methods include Principal Component Analysis (PCA), Multidimensional Scaling (MDS), t-Distributed Stochastic Neighbor Embedding (t-SNE), and others.

The first algorithm, PCA, involves reducing linear dimensionality by transforming data into a new coordinate system, making it easy to identify directions (principal components) that reflect the most significant data variation. It is utilized to decompose a multidimensional dataset into a series of sequential orthogonal components that explain the maximum variance. In PCA, the first principal component of a set of variables is a derived variable formed as a linear combination of the original variables, explaining the most variance. The second principal component explains the most variance of what remains after eliminating the effect of the first principal component, and so forth.

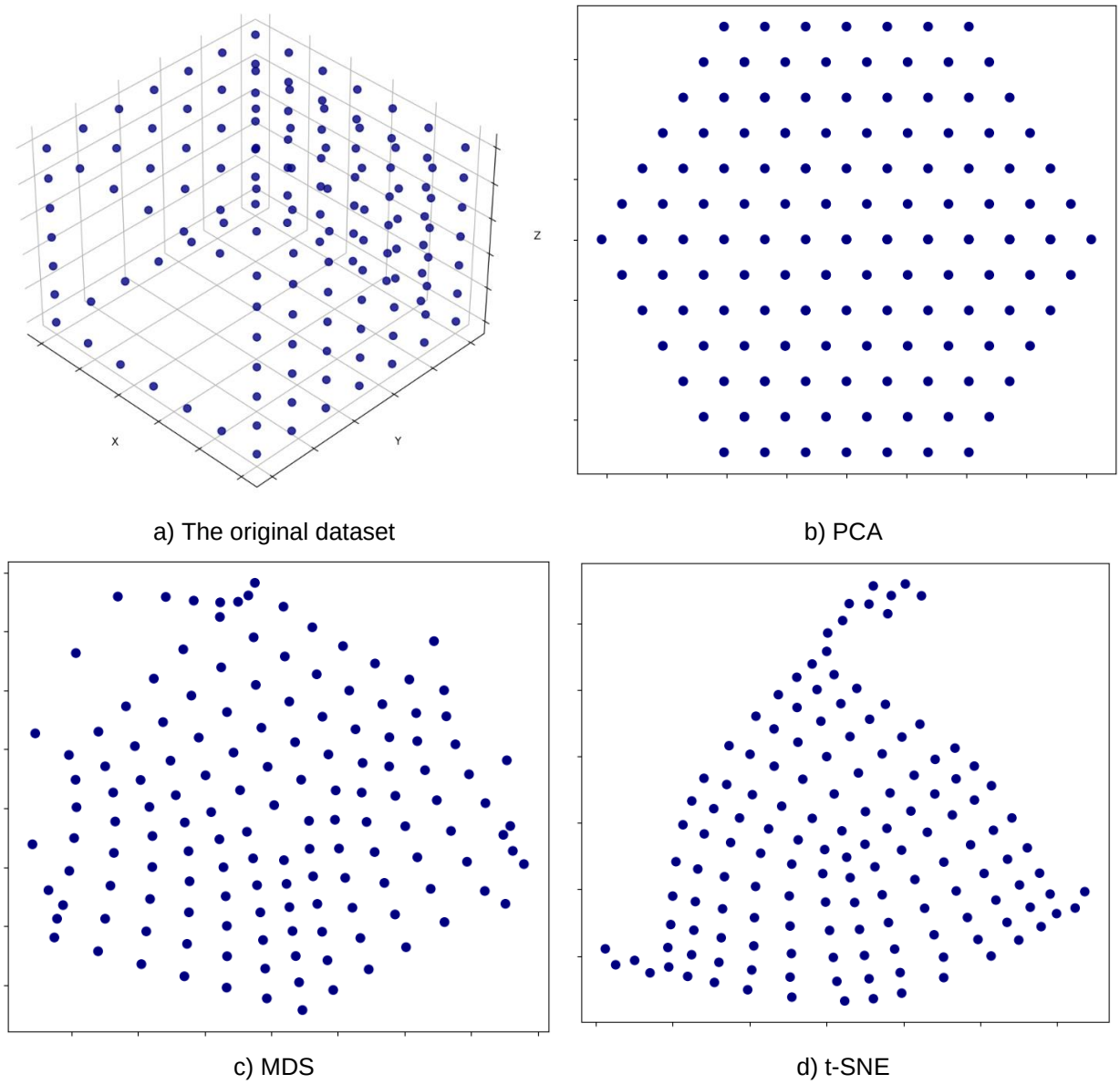
Another algorithm, MDS, seeks a low-dimensional representation of data where distances accurately reflect those in the original high-dimensional space [7]. MDS positions each object in a lower-dimensional space in such a way that distances between objects are preserved as faithfully as possible. This method constitutes a type of non-linear dimensionality reduction.

t-SNE is a method of non-linear dimensionality reduction that represents each high-dimensional object with a two- or three-dimensional point in such a way that similar objects are modeled by nearby points, while dissimilar objects are modeled by distant points with high probability [8]. The t-SNE algorithm consists of two main stages. Firstly, t-SNE constructs a probability distribution over pairs of high-dimensional objects such that similar objects are assigned higher probabilities, while dissimilar ones are assigned lower probabilities. Secondly, t-SNE determines a similar probability distribution over points in the low-dimensional space and minimizes the Kullback-Leibler divergence between the two distributions.

The dimensionality reduction methods are effective as they can decrease the number of dimensions required to describe data while retaining their internal structure. However, two challenges accompany the dimensionality reduction methods:

1. They compress a large number of dimensions into fewer ones. Consequently, data points that are well separated in the multidimensional space may become neighbors or even merge into a single point in the projection. This increases the likelihood of missing crucial information.
2. Different dimensionality reduction algorithms yield different results. None of these results are incorrect; they simply differ because various algorithms emphasize different aspects of the data. Relying on a single algorithm may lead to the oversight of important information.

Figure 2 illustrates an example of dimensionality reduction using the aforementioned methods from a three-dimensional space to a two-dimensional space.



a) The original dataset

b) PCA

c) MDS

d) t-SNE

Figure 2. An example of dimensionality reduction methods: a) The original dataset is represented by a cube with three of its faces left unshaded; b) PCA projection; c) MDS projection; d) t-SNE projection

Figure 3 presents a graph constructed from the dataset. The graph does not exhibit the typical issue encountered in the dimensionality reduction methods, where points that are distant in a high-dimensional space may overlap in a low-dimensional space.

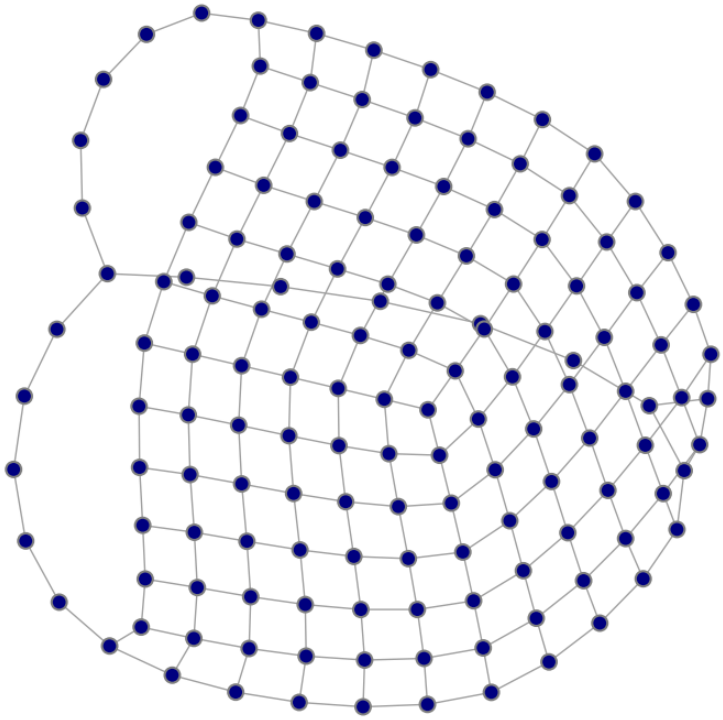
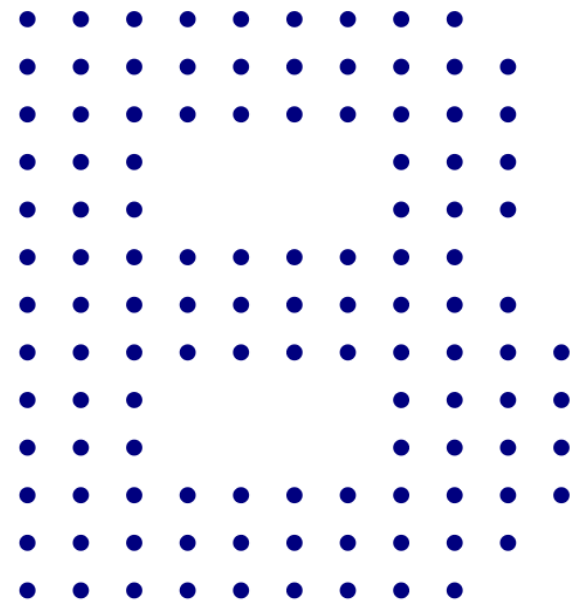


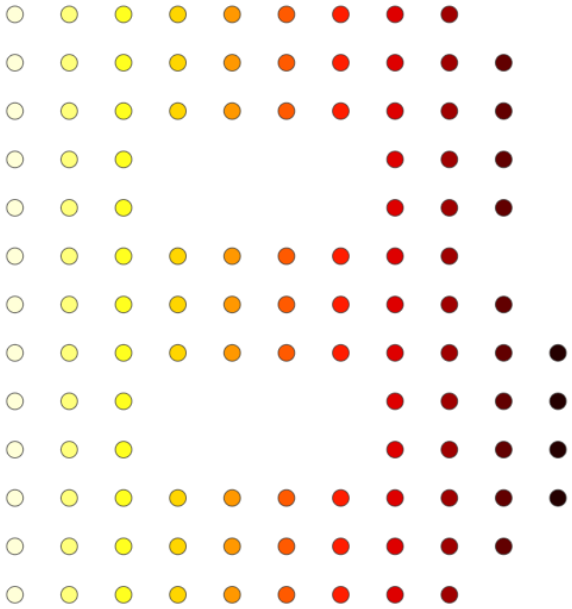
Figure 3. Graph built from the dataset

To construct the graph, several steps of the algorithm need to be followed. Let us illustrate this with a two-dimensional dataset shown in **Figure 4 a**:

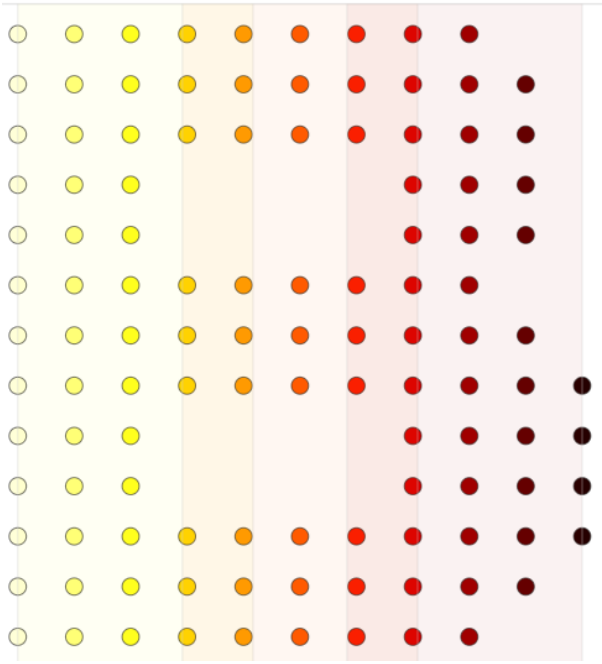
1. We define a function, called a projection function, on the dataset points, in our case we perform the projection onto the x-axis. In **Figure 4 b**, the points are colored according to the projection values.
2. Next, we build a covering. Specifically, we order the data points according to their projection values and divide them into overlapping intervals (**Figure 4 c**).
3. We then construct the graph, where the nodes of the graph are the data points, and an edge between two nodes is drawn if these nodes lie within the same interval and are close according to the chosen metric (**Figure 4 d**). Since the graph is a mathematical object consisting of nodes and edges, to visualize it, we need to calculate the coordinates of the nodes to position them on the plane of the figure. The graph's orientation differs from that of the original dataset because the nodes positions are calculated without knowledge of the coordinates of the points in the dataset.



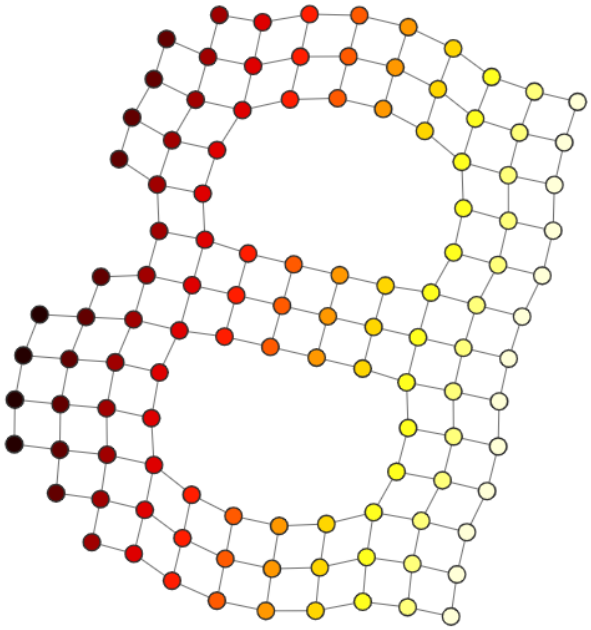
a) The original dataset



b) The original dataset, colored by the projection



c) A covering with overlapping intervals



d) The graph

Figure 4. Stages of constructing the graph: a) Original dataset; b) Dataset with points colored by projection; c) Coverage, i.e., division into overlapping intervals; d) The resulting graph

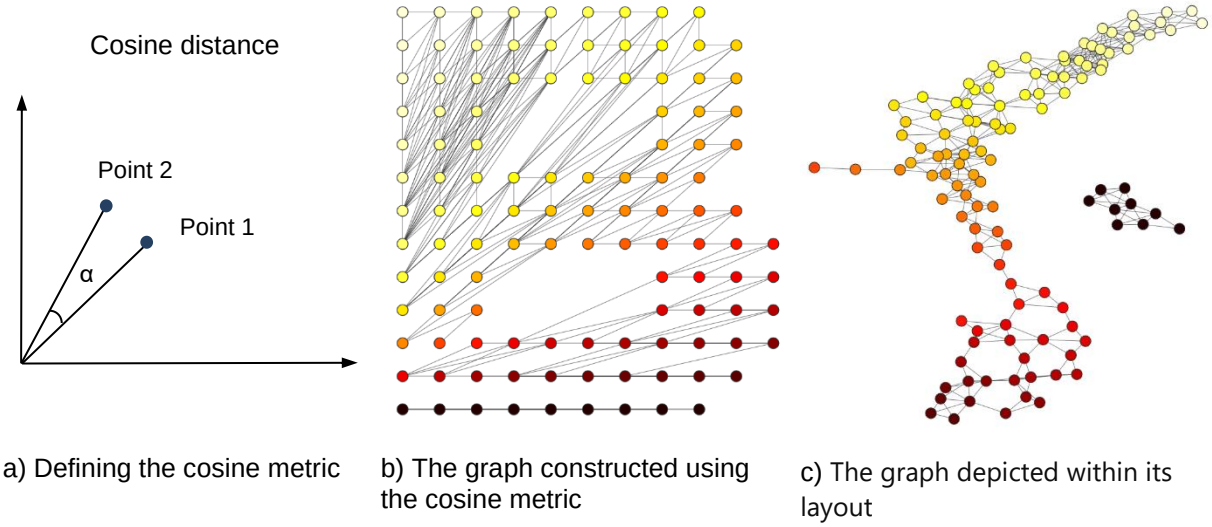
The proximity of points in the dataset, which determines the presence of edges in the graph, is defined by a metric (a distance function).

Distance metrics are functions $d(a, b)$ such that $d(a, b) < d(a, c)$ if objects a and b are considered "more similar" than objects a and c. Two identical objects have a distance of zero. One of the most popular examples is the Euclidean distance, which was used in the previous example. To be a "true" metric, it must adhere to the following four conditions:

1. non-negativeness: $d(a, b) \geq 0$, for all a and b;
2. positive definiteness: $d(a, b) = 0$ if and only if $a = b$;
3. symmetry: $d(a, b) = d(b, a)$;
4. the triangle inequality: $d(a, c) \leq d(a, b) + d(b, c)$.

A graph based on a two-dimensional dataset can be constructed using a different metric, such as the cosine metric, which represents the cosine of the angle between points denoted as vectors as shown in **Figure 5** a. The angle is what matters, irrespective of how far the points are from the origin, or how large their coordinate values are. In the graph, edges connect vertices that form a small angle between them. The nodes of the graph can be plotted using the coordinates of the dataset points, as depicted in **Figure 5** b, or the graph layout can be computed (see **Figure 5** c).

The selection of the metric for the graph depends on the specific task at hand. It's vital to take this into account when comparing the similarity of two patients; modifying the metric might result in a different graph.



a) Defining the cosine metric

b) The graph constructed using the cosine metric

c) The graph depicted within its layout

Figure 5. Constructing a graph using the cosine metric: a) Defining the cosine metric; b) The graph constructed using the cosine metric, with nodes coloring based on the angle relative to the x-axis; c) The graph whose nodes are positioned according to the coordinates of the layout

When calculating the covering, we can divide the projection into equally overlapping intervals, which we refer to as a uniform coverage. It is convenient to use the uniform coverage if the projection values are evenly distributed across the entire range. In the case of uneven distribution, a balanced coverage is preferable, in which each interval contains the same number of points.

Multiple projection functions can be used, and then the multidimensional projection is obtained as a combination of several one-dimensional (uniform or balanced) projections. The combination is obtained using the following algorithm: by iterating through all possible pairs of intervals from two projections, we select points belonging to both intervals simultaneously (the intersection of intervals).

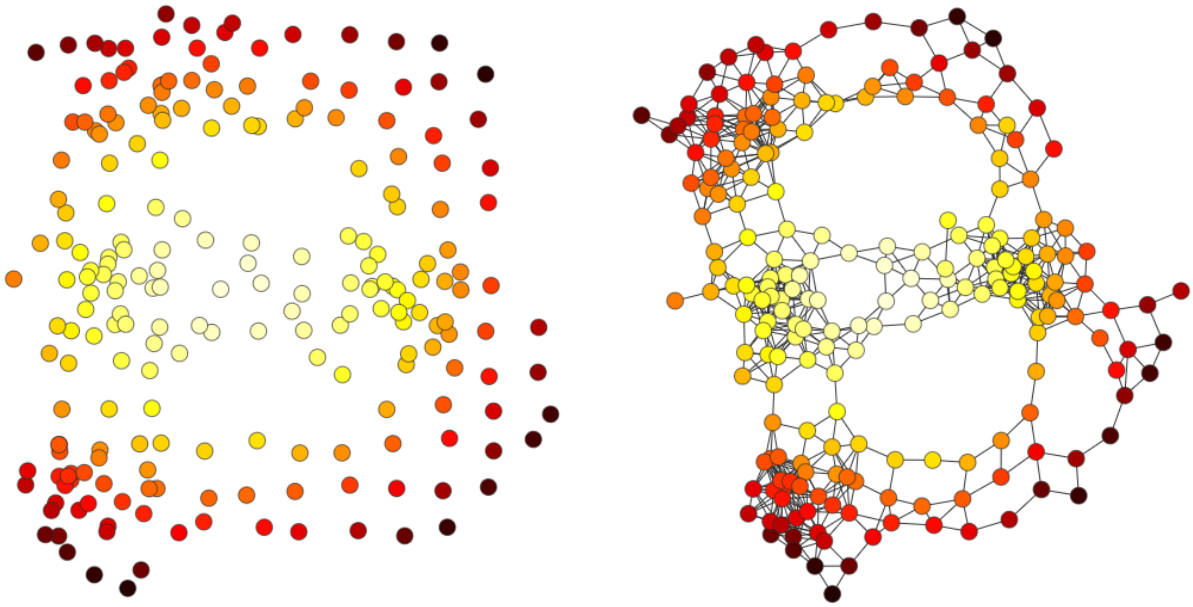
In the case of a two-dimensional covering, each interval represents a rectangle, and the covering consists of a set of overlapping rectangles. In the three-dimensional case, we have intersecting parallelograms.

Let us explore some other functions for projections.

The Centrality projection indicates the distance of a point from the 'center' of the data or how well a point conforms to the “norm” versus being an outlier. This function has one parameter p, and is computed using the following formula:

$$Centrality(x) = \begin{cases} \left(\frac{\sum_{i=1}^{N} d(x,x_i)^p}{N}\right)^{1/p}, \ if\ 1 \le p < +\infty \\ \max\limits_{i} d(x,x_i), \ if\ p = +\infty \end{cases},$$

where d is a distance function, N is a number of points.



a) Original dataset, colored according to the projection

b) Resulting graph

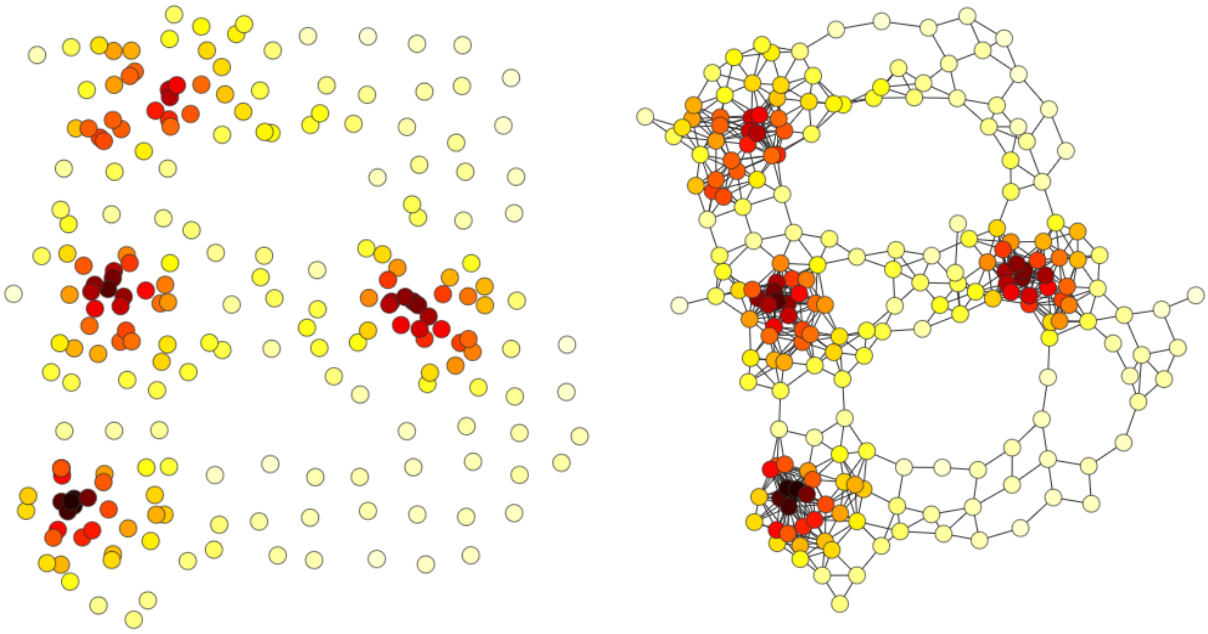
Figure 6. A graph with the Centrality projection: a) Original dataset, colored according to the projection; b) Resulting graph

The Density projection assesses the density of neighboring points around a given point and is calculated using the following formula:

$$Gauss\_density(x) = \frac{1}{N(\sqrt{2\pi}\sigma)^n} \sum_{i=1}^{N} \exp\left(\frac{-d(x,x_i)^2}{2\sigma^2}\right),$$

where σ^2 is a scale parameter.

The graph with the Density projection is shown in **Figure 7**.



a) Original dataset, colored according to the projection

b) Resulting graph

Figure 7. A graph with the Density projection: a) Original dataset, colored according to the projection; b) Resulting graph

We can employ both the previously discussed dimensionality reduction methods and statistical functions such as Max, Mean, Median, Variance, Entropy, and others as projections. Data-driven projections involve variables from the data that were not utilized in constructing the original dataset. For example, in clinical research, patient's age may serve as a projection. In this case, the graph would represent patient stratification by age, connecting patients not only when they are close according to the chosen metric but also when they are similar in age.

After the graph is constructed according to selected variables, the researcher then visually explores it to discover interesting subgroups within the data. For example, the isolated components of the graph or highly interlinked groups of nodes that form communities may indicate meaningful relationships within the dataset.

2.3. AUTOMATIC COMMUNITY DETECTION

The key feature of a graph is a community structure, which relates to the way the nodes are organized in communities. Specifically, many edges connect nodes within the same community (or cluster), while comparably few edges connect nodes between different communities. These clusters or communities can be considered to represent independent structures within the graph, and the detection of those independent communities is one of the key goals in the analysis of large graphs that represent complex relationships within datasets.

In graphs that represent real-world systems, the distribution of edges over subgroups of nodes is usually non-uniform. This reflects the possible presence of some hidden structures and patterns in the graph, and hence in the real-world data from which the graph was created. Specifically, some groups of nodes may have high concentrations of edges, while the concentrations of edges between these groups of nodes may be low. This structure takes the form of an intermediate-scale graph structure known as a community structure, or a cluster structure, where a group of densely connected nodes is referred to as a community. **Figure 8** illustrates an example of a community structure within a graph that contains three clusters of nodes with dense internal connections and comparably fewer connections between clusters.

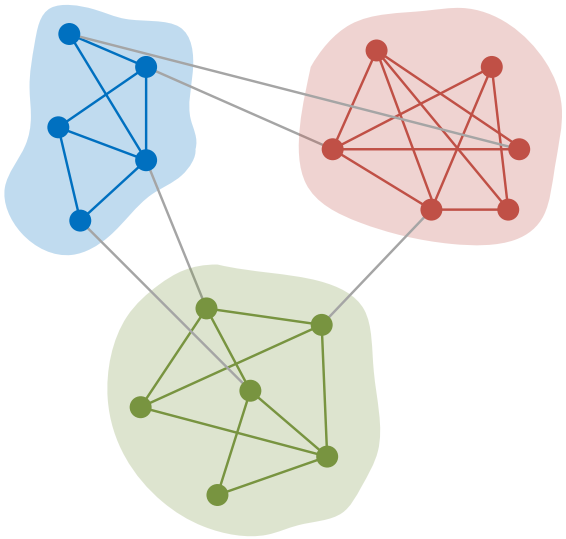


Figure 8. A schematic representation of a simple graph that has a community structure. The graph contains three communities of densely connected nodes that have a much lower density of connections (gray edges) between them

Communities, or clusters, are groups of nodes within a graph that are likely to share common properties and/or play similar roles. In view of this, where possible, the aim of community detection is to identify communities within the graph and their hierarchical organization by using the information that is contained within the graph topology alone. Identifying the communities according to the topological properties of the graph only allows the classification of nodes according to their structural position on the graph. Thus, nodes with a central position in their communities share the largest number of edges with the other nodes within the community, which may indicate the important role they play in the stability of the community. On the other hand, nodes that are located at the boundaries between communities may play an important role as mediators in the relationships and exchange between different communities.

The problem of graph clustering, intuitive at first sight, is actually not well defined. Though numerous attempts have been made to analyze real-world systems based on the community structure in multiple disciplines and through practical applications, graph theory does not define the problem of graph clustering, and no universally accepted definitions of a community or partitioning into communities have emerged. Therefore, the concepts of a community and partitioning into communities are to some extent arbitrary and must be determined by researchers according to the specific problem under consideration.

Detecting communities within a graph (especially large ones) can be computationally difficult if the number of communities within the graph is unknown and the size and density of the communities are unequal.

2.4. A WORKFLOW FOR GRAPH-BASED DATA ANALYSIS

TDA is used to create a flexible and versatile workflow to perform graph-based data analysis. This workflow can be adapted to a variety of scenarios and types of data in order to identify hidden patterns. The key steps are summarized and highlighted in **Figure 9**. We see their implementation in our experiment in Section 3. All of the steps in the workflow except Steps 1, 5, 8, are performed automatically using ML algorithms.

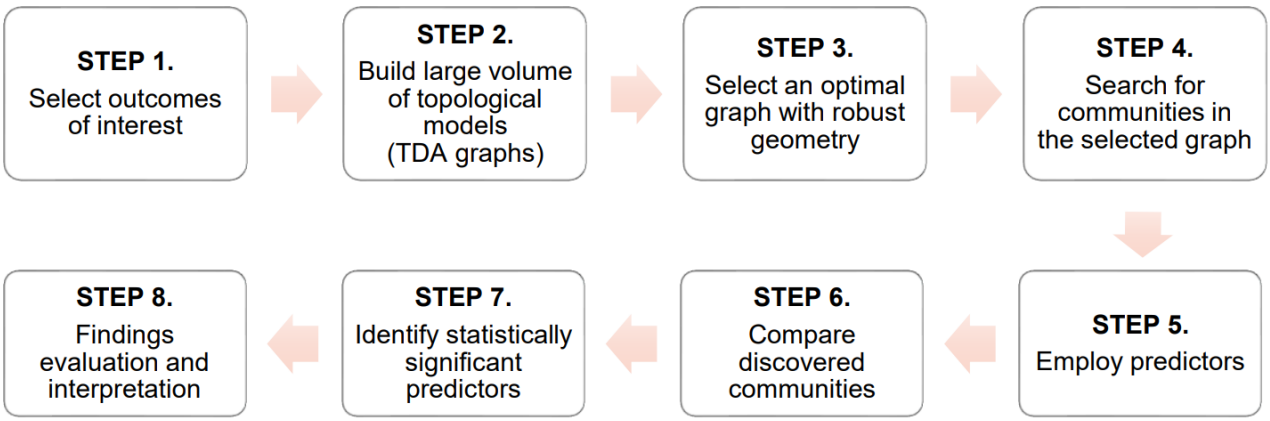


Figure 9. Workflow of graph-based data analysis

Let us expand on the steps in our workflow:

STEP 1. From a given dataset, the outcomes of interest are selected. At this step, some pre-processing of data might be required to deal with irregularity, e.g., to account for missing data, to aggregate noisy data, etc.

STEP 2. Using the selected outcomes of interest, a large volume of TDA-graphs is built by varying parameters of the TDA algorithm, e.g., the parameters in the distance function, in the projection function, etc. (see Section 2.2).

STEP 3. The most robust and representative graph is selected based on an array of criteria, e.g., an adapter modularity score, Kolmogorov complexity, etc. In many applications, the majority vote or the cumulative ranking among the optimality scores selects the most representative graph.

STEP 4. A selection of community detection algorithms is applied to the most representative graph at this step to reveal hidden patterns within data in form of communities on the graph. The discovered communities are highlighted on the graph by coloring and are subject to further analysis.

STEP 5. A selection of predictors of interest is integrated into the model to explain the detected communities, and hence to explain hidden similarities within the dataset of study. At any time, additional predictors can be incorporated into the model at this step to expand the search of unrelated features.

STEP 6. Communities on the graph correspond to subsets of patients. A comparison of communities is performed at this step, e.g., by comparing sizes, overlap, persistence over different community detection methods, etc.

STEP 7. Further pairwise or community-against-the-rest comparison of communities is done at this step using statistical analysis based on predictors. Statistically significant predictors are selected. This step helps to identify the key variables that are driving the community structure and involves a large volume of automatic statistical tests.

STEP 8. At this final step, the statistically significant predictors of the discovered community structure are being further interpreted, e.g., using subject-matter expertise.

3. KRI VISUALIZATION WITH TDA

In this section, we follow our workflow (see **Figure 9**) to demonstrate some of visual discovery tools that TDA can provide for analysis of a KRI focusing on a patient retention problem. We consider patients who discontinued treatment and those subjects who discontinued the study as KRIs of interest (for more KRIs refer [9]). The reasons for the subject discontinuation may vary from a serious adverse event to quitting a study without a known reason. Thus, absent records of visits or empty data are also included to this KRI.

The analysis was performed on all the 582 patients of CSP1018 study [3]. Having selected the outcomes of interest, all missing values were filled with zeros following the aim to track “zero”-patients (either true or

imputed) which are interpreted as subjects who discontinued treatment or the study. After building a topological model, we detect and study groups of patients (communities) with similar data patterns and discover significant differences between predictors, while paying special attention to “zero”-communities.

We build several topological models that visualize the data from different perspectives. The models demonstrate how different parameters (metrics and projection) and outcomes of interest affect visual representation of the same data in the form of a graph. Every node of each graph corresponds to a patient and two nodes are connected if they are similar in outcomes based on the selected distance and projection metrics. Focusing on the geometric properties of datasets, the experiment aims to unveil a set of unrelated features that could have caused similarities in the discontinuation KRI. A representative graph automatically highlights possible similarities within the data and discovers meaningful subgroups of patients as communities on the graph. Further, by integrating various predictors from different areas, such as demographic, social, etc., the revealed similarities can be described.

In this section, we demonstrate our TDA workflow on two classes of graphs, namely 1) ~~a~~ graphs built on prescribed milligrams (mgs) of buprenorphine outcome, and 2) a graph built on a visit rates outcome.

3.1. DATA EXPLORATION BY A PRESCRIBED MILLIGRAMS OF BUPRENORPHINE OUTCOME

In this subsection, we consider prescribed mgs of buprenorphine therapy taken as an outcome of interest. After preprocessing the data, each subject is assigned a 532-dimensional row-vector with prescribed buprenorphine mgs per day values (see **Figure 10**) corresponding to the 532-day duration of the whole period of study. This vector represents different dynamics of ongoing treatment of the subject.

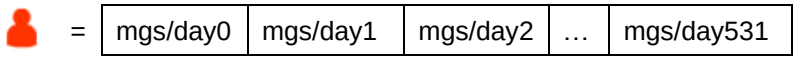


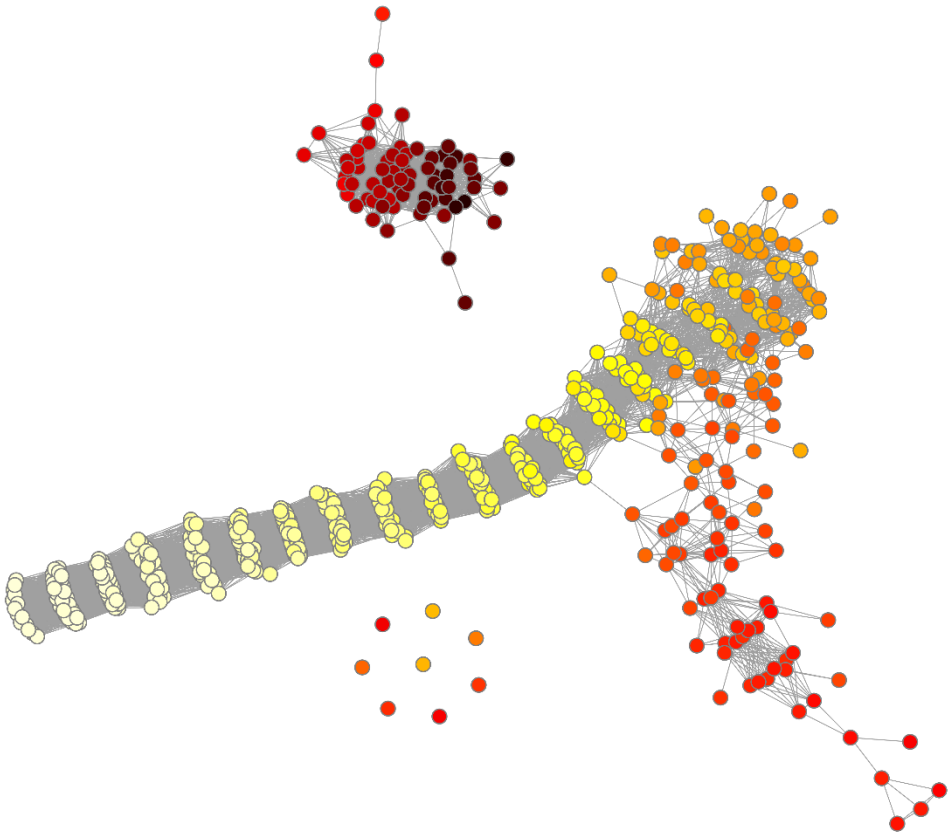
Figure 10. **A 532-dimensional row-vector with daily values of buprenorphine mg prescribed**

Patients that have zero values starting from an early study day are patients with a high KRI of *treatment discontinuation* and are the focus of our interest in this subsection.

Building different topological models

Let us explore different TDA models built on these data vectors in different projections and distances metrics. Each node of the following graphs represents a single patient with the total of 582 nodes.

Figure 11 visualizes data by mgs in density projection and l1 Manhattan distance metrics, which is the sum of absolute differences between vectors’ coordinates across all the dimensions. The coloring is performed by mean mgs values prescribed during the whole study period varying from light yellow (small mgs values) to dark red (large mgs values).



Legend: *less mgs* - - *more mgs*

Figure 11. TDA graph in *density projection* and *l1 Manhattan distance metrics* of the *mgs per day* dataset. The coloring reflects the mean mgs values prescribed during the whole study period

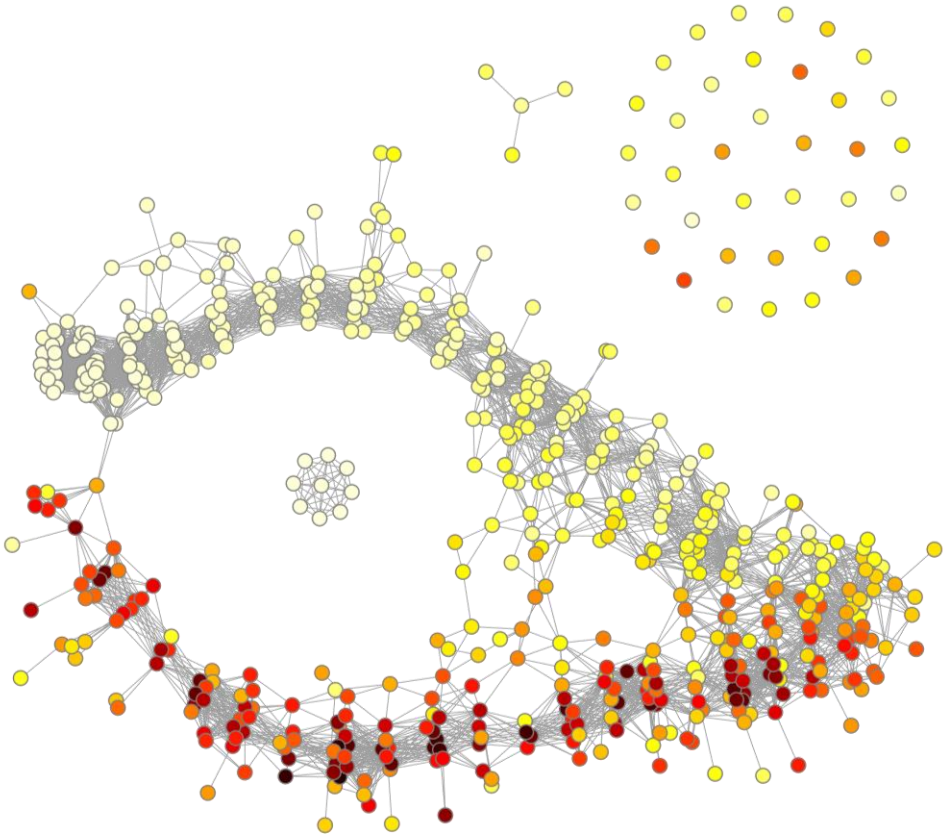
Let us consider a **correlation distance**. It is computed as follows:

$$d_c(u,v) = 1 - \frac{(u-\bar{u})\cdot(v-\bar{v})}{\|(u-\bar{u})\|_2\|(v-\bar{v})\|_2},$$

where $\bar{v}$ is the mean values of the elements of a vector v, $\|\ldots\|_2$ is the standard Euclidean distance, and $a \cdot b$ is the dot product of vectors a and b.

The correlation distance ranges from 0 to 2 and measures proximity of directions of trends. Thus, it is sensitive to the slopes of the time series. However, it is centered, so it is not sensitive to the shifts and does not distinguish the starting points of the trends (the baseline values of prescribed mgs) and thus the mean mgs values during the visits.

Building a graph using the correlation metric (see **Figure 12**) incorporates similarity in the dynamics of treatment doses but does not capture the absolute amount of buprenorphine prescribed. The graph observed in **Figure 12** is built using the centrality projection and nodes with similar trends are connected with edges. The coloring is identical to that in **Figure 11** and is performed by mean mgs values.



Legend: *less mgs* - - *more mgs*

Figure 12. The TDA graph build using the *centrality projection* and *correlation metrics* of the *mgs per day* dataset. The coloring reflects the mean mgs values prescribed during the whole study period.

Finally, let us introduce a **combined metric:**

$$d_{cm}(\text{♂}, \text{♂}) = d_c(\text{♂}, \text{♂}) + d_m(\text{♂}, \text{♂}),$$

which is used to measure the dissimilarity between participants. The metric d_{cm} is composed of the correlation distance, which was introduced above, and the following **absolute value of difference between the mgs means:**

$$d_m(u, v) = 2\left|\frac{\bar{u} - \bar{v}}{\max_{a,i} a_i}\right|,$$

where max is taken over all coordinates a_i of all vectors a of the dataset (i.e., it is maximum mgs values observed). Means are normalized in such a way that d_m ranges from 0 to 2. It is done in order to balance d_m contribution with the one provided by the correlation distance d_c.

This metric is designed to capture the difference in treatment trends (increasing or decreasing) as well as the mean values of mgs of buprenorphine prescribed during the whole period of the experiment. The corresponding graph built using the centrality projection is presented in **Figure 13**.

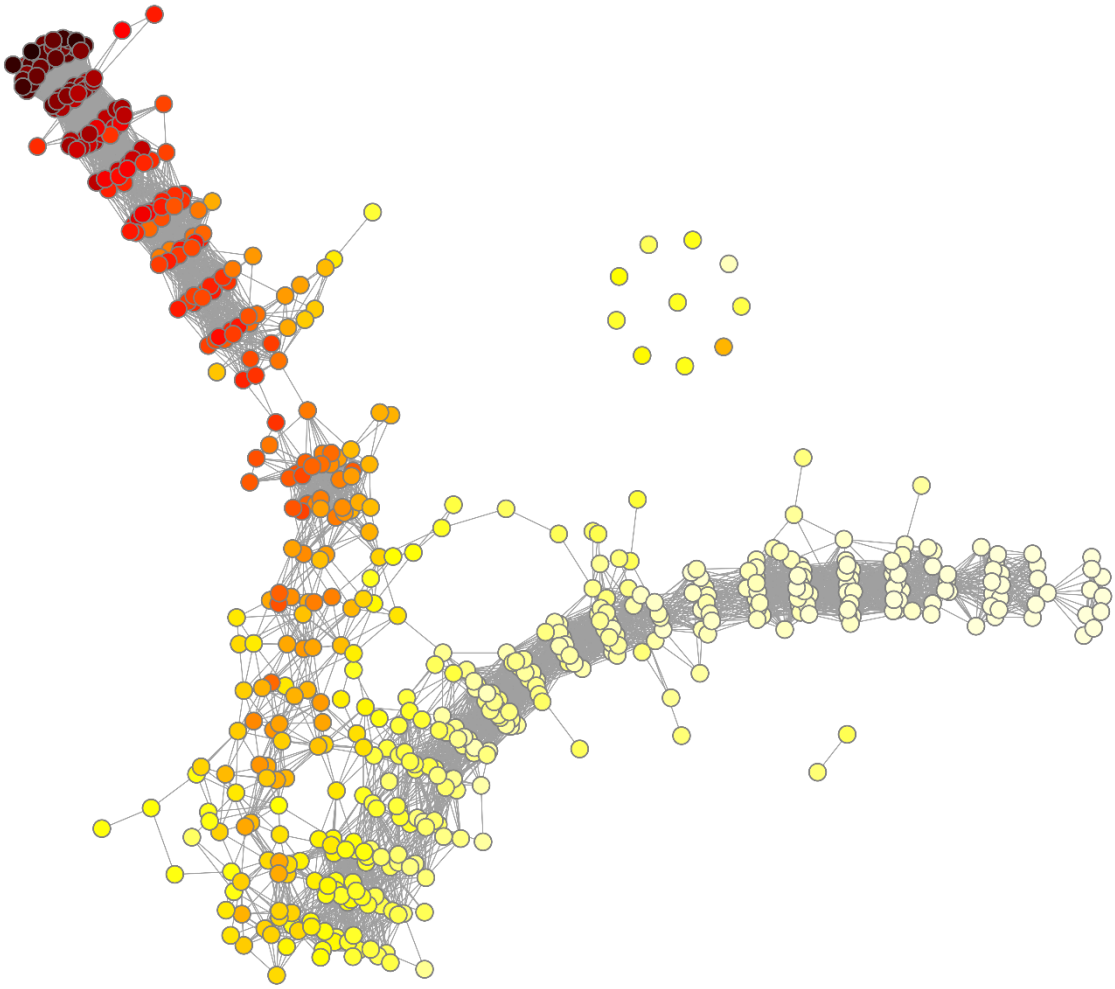


Figure 13. The TDA graph built using the *centrality projection* and *combined metrics* of the *mgs per day* dataset. The coloring reflects the mean mgs values prescribed during the whole study period

Community detection and description

After the topological model is extracted, a community detection algorithm is applied to reveal interesting subgroups within the data. Let us demonstrate results of the clique percolation algorithm with a 5-clique applied to the last graph. The discovered communities are highlighted on the graph by coloring (**Figure 14**). The method finds 7 communities and 75 non-community gray nodes. Note that the smallest *third* (pink) community is comprised of patients with no data at all (quitted the study at the very beginning).

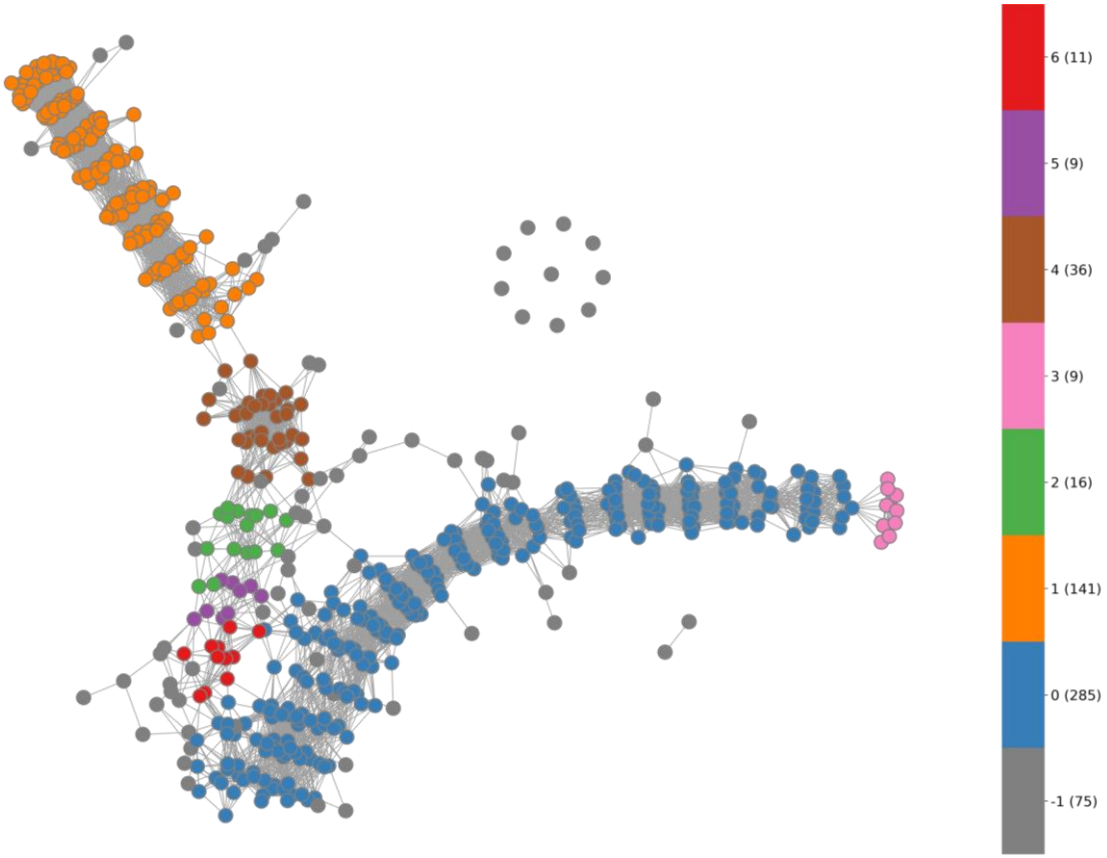


Figure 14. The output of the 5-clique percolation method

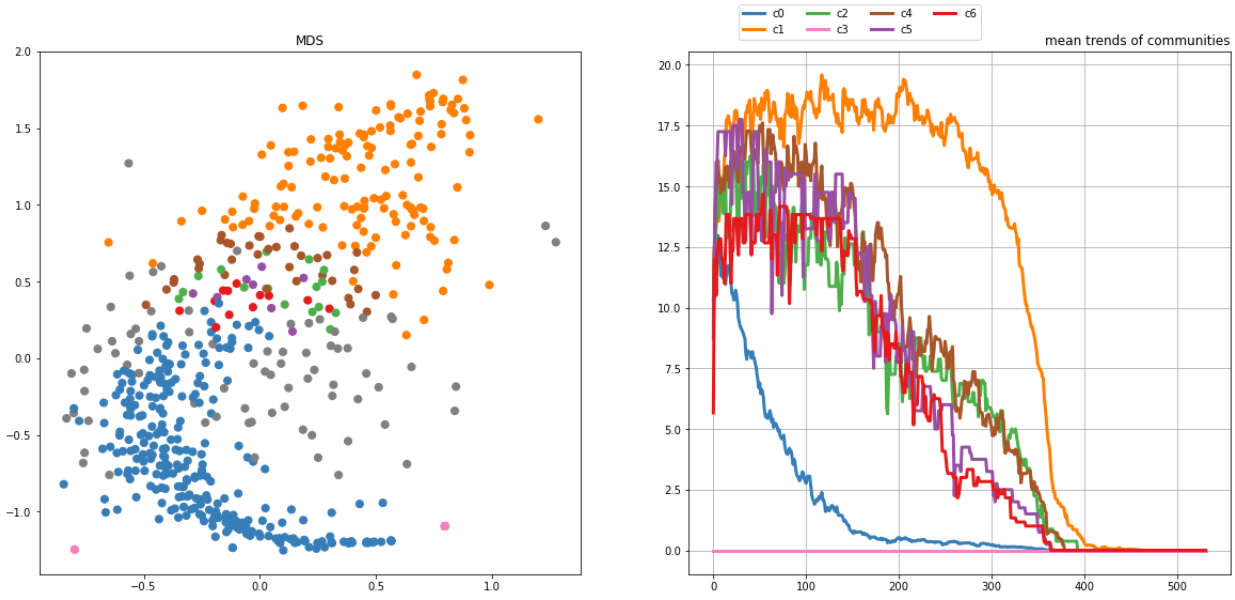


Figure 15. The MDS projection and mean trends of communities

Figure 15 indicates that MDS projection is well aligned with the TDA graph, although some points may be invisible in the MDS projection due to the overlay. The mean trends of the percolation communities well stratified by the outcomes.

Let us incorporate various predictors in order to explain the patterns found by the clique percolation method.

Running statistical tests, a table of statistically significant predictors with the p-values < 0.05 (a level of significance) is calculated to establish that the distribution of the predictors for a selected community of nodes was different from that of another community or all the rest of the dataset (i.e., complementary points). After a significance level of any predictor was found to be statistically significant, we are able to construct a histogram representing normalized frequency distributions of the predictor for both the nodes in the selected pair of communities or for the community and its complementary.

For the purpose of the statistical analysis, continuous, mixed, binary, and categorical (non-binary) univariate predictors were differentiated according to a variable type. Continuous predictors were examined using the standard two-sample Mann–Whitney–Wilcoxon test. To examine the statistical association between two samples within the categorical data, the Fisher's exact test and the χ^2-square test were used for the binary and non-binary categorical variables, respectively.

We focus our main attention on the two communities with low retention rates (a high treatment discontinuation KRI). These communities are the **third (pink)** community which comprises absolutely zero values (i.e., patients having no visits and completely empty mgs data) and the **zeroth (blue)** community with rapidly decreasing to zero mgs rates (it is the biggest community) (see **Figure 14**). Several significant differences (p-values < 0.05) are presented below. We do not set ourselves the task of providing a comprehensive analysis and simply show a few examples.

The **zeroth (blue)** community has a significantly higher rates of 'will to work with study physician', lower values of alcohol dependences, days outpatient and heroine in past 30 days values (see **Figure 16**).

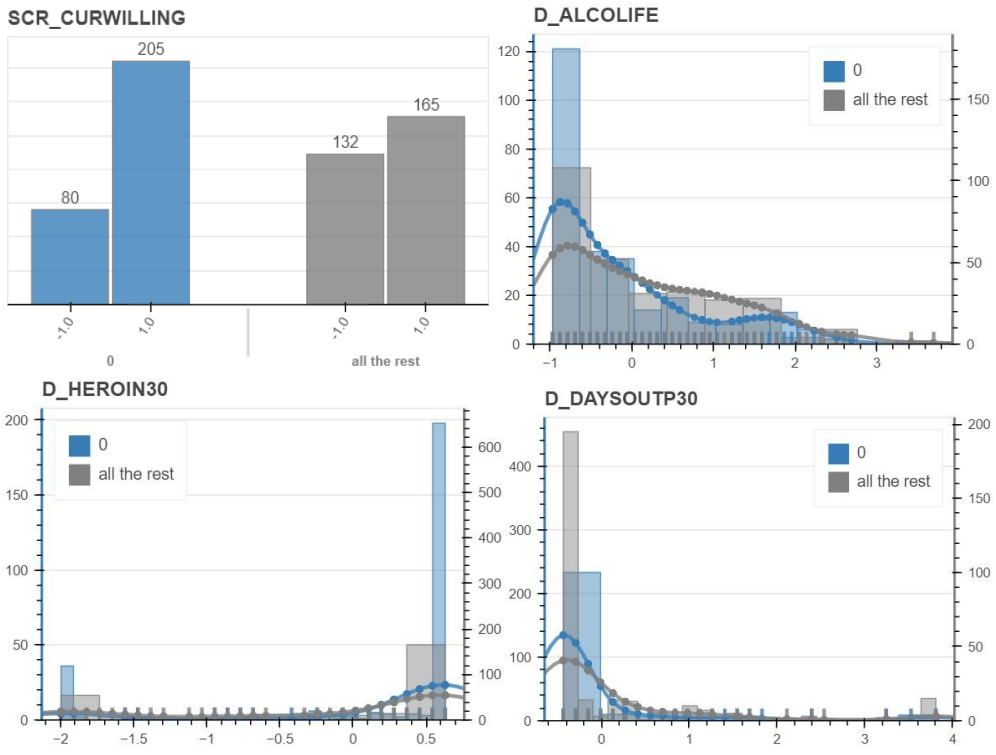


Figure 16. The zeroth (blue) community vs all the rest of the dataset

As to the **third (pink)** community, it differs significantly with all the rest of the dataset by the predictor of legal problems with prostitution. As to the pairwise comparison, it shows significant differences in race distribution with the **first (orange)** community with the lowest treatment discontinuation KRI (see **Figure 17**).

Comparing the **first (orange)** community with other communities, a significant difference can also be observed by heroine in past 30 days and socio-employment pattern predictor (**Figure 18**).

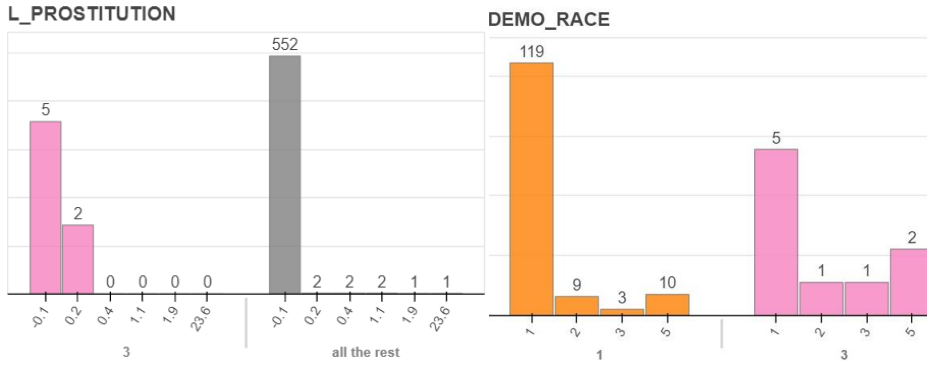


Figure 17. The third (pink) community

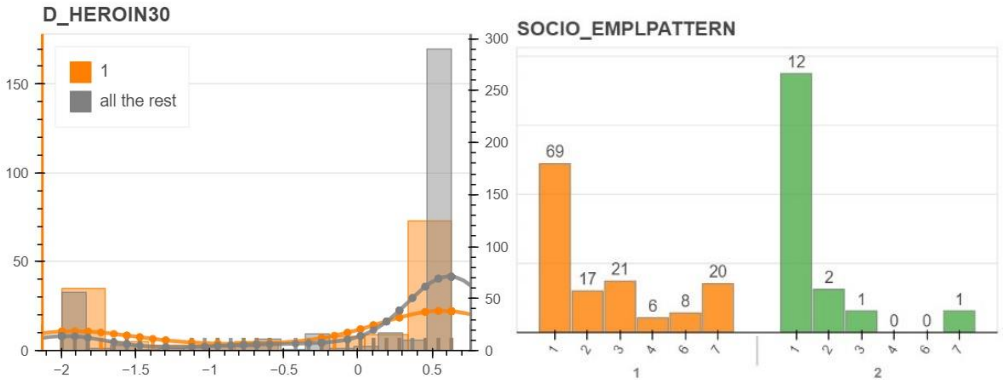


Figure 18. The first (orange) community

3.2. DATA EXPLORATION BY A VISIT RATES OUTCOME

In this subsection, we do not focus on treatment discontinuation as a KRI. Instead of treatment outcomes, we consider visit rates (VR) as outcomes, which allows us to capture the KRI of *subject discontinuation*.

Let us describe in more detail how the VR values are constructed. Following the protocol (see [3]), the whole period of study is partitioned into 6 parts: week 1, weeks 2-6, weeks 7-12, weeks 13-26, weeks 27-52, and weeks 53-76. Each part contains a fixed number of expected visits: 2 visits during the first week, 5 visits weekly during weeks 2-6, 6 visits weekly during weeks 7-12, 7 visits once in two weeks during weeks 13-26, 6.5 visits once in four weeks during weeks 27-52, and 6 visits once in four weeks during weeks 53-76. We also incorporate into the VR values taper and complement data starting from week 7 (according to the protocol) by adding 1 in the numerator for the taper and assigning plus 1 to the rate if the patient successfully completed the study. The corresponding formulas for VR values are the following:

VR1 = (# of visits during week 1)/2,

VR2 = (# of visits during weeks 2-6)/5,

VR3 = ((# of visits during weeks 7-12) + taper)/6 + complete,

VR4 = ((# of visits during weeks 13-26) + taper)/7 + complete,

VR5 = ((# of visits during weeks 13-26) + taper)/6.5 + complete,

VR6 = ((# of visits during weeks 53-76) + taper)/6 + complete.

Thus, each subject is assigned a 6-dimensional row-vector with VR values (see **Figure 19**) for 6 different periods of the study. This vector represents the attendance dynamics, retention, and successful adherence to protocol.

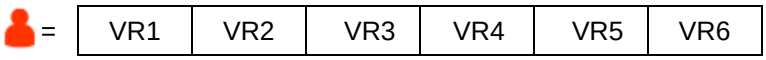


Figure 19. The 6-dimensional row-vector with visit rates during 6 parts of the whole period of the study

Building a topological model

Let us now explore a TDA model built on these data vectors. **Figure 20** visualizes data by visit rates in the centrality projection and l1 Manhattan distance metrics. The coloring corresponds to that in **Figure 11**, **Figure 12**, and **Figure 13** and is performed by mean mgs values prescribed during the whole study period varying from light yellow (small mgs values) to dark red (large mgs values).

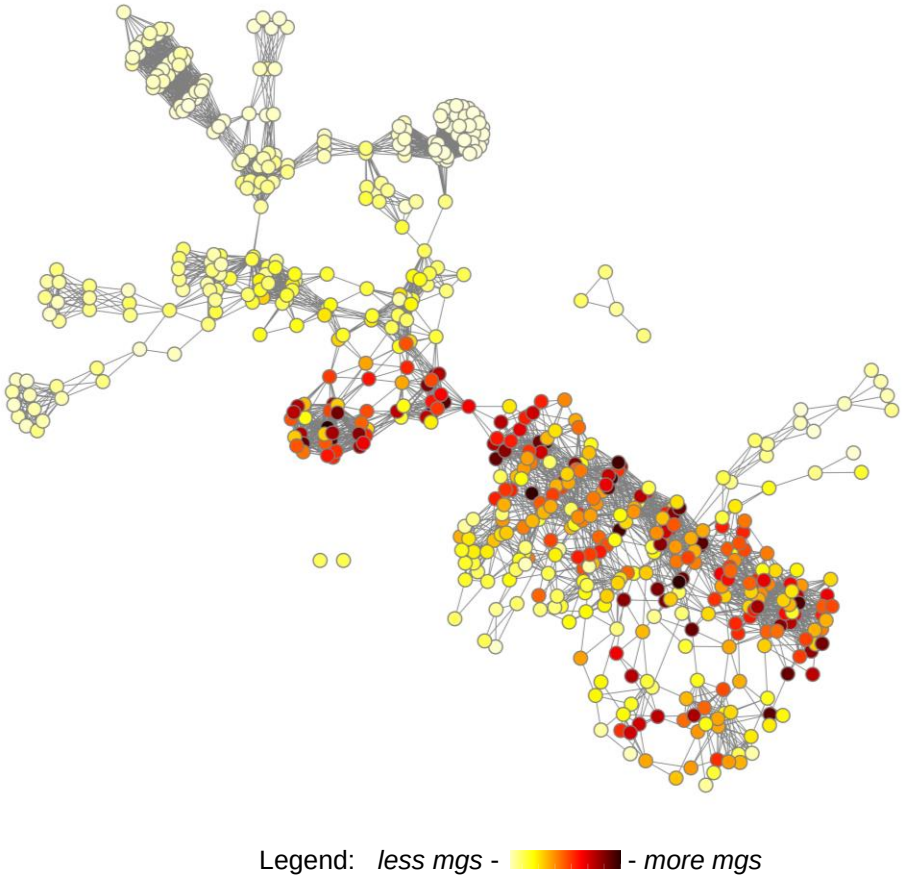
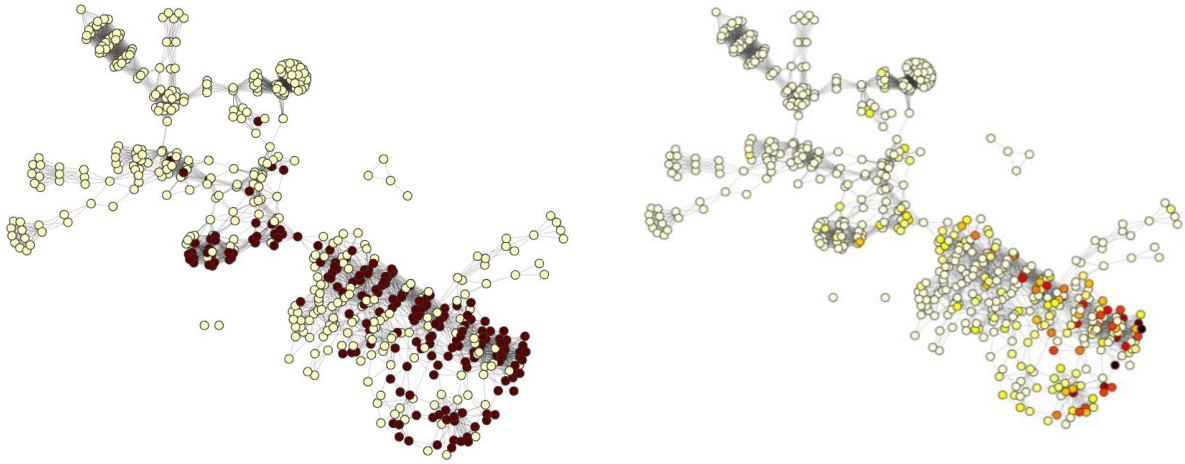


Figure 20. **The TDA graph built using the *centrality projection* and *l1 Manhattan distance metrics* based on the *visit rates* dataset. The coloring reflects the mean mgs values prescribed during the whole study period**

We can highlight patients of interest by using different colorings. For instance, coloring nodes by completion status or prescribed buprenorphine tapering of patients (**Figure 21**).



a) by completion status (light yellow – not completed, dark red - completed)

b) by taper rates (*less* - - *more*)

Figure 21. **Coloring of the TDA graph**

Communities detection and description

Upon applying the 4-clique percolation algorithm (**Figure 22**), we obtain 16 communities and 39 non-community grey nodes. As it can be easily observed, the method distinguishes communities (the biggest *second* and *first* ones as well as the *fourth* and *fifteenth* small ones) which mostly comprise patients who had completed the study (with a low KRI of subject discontinuation) and all the other communities (with a high KRI of subject discontinuation).

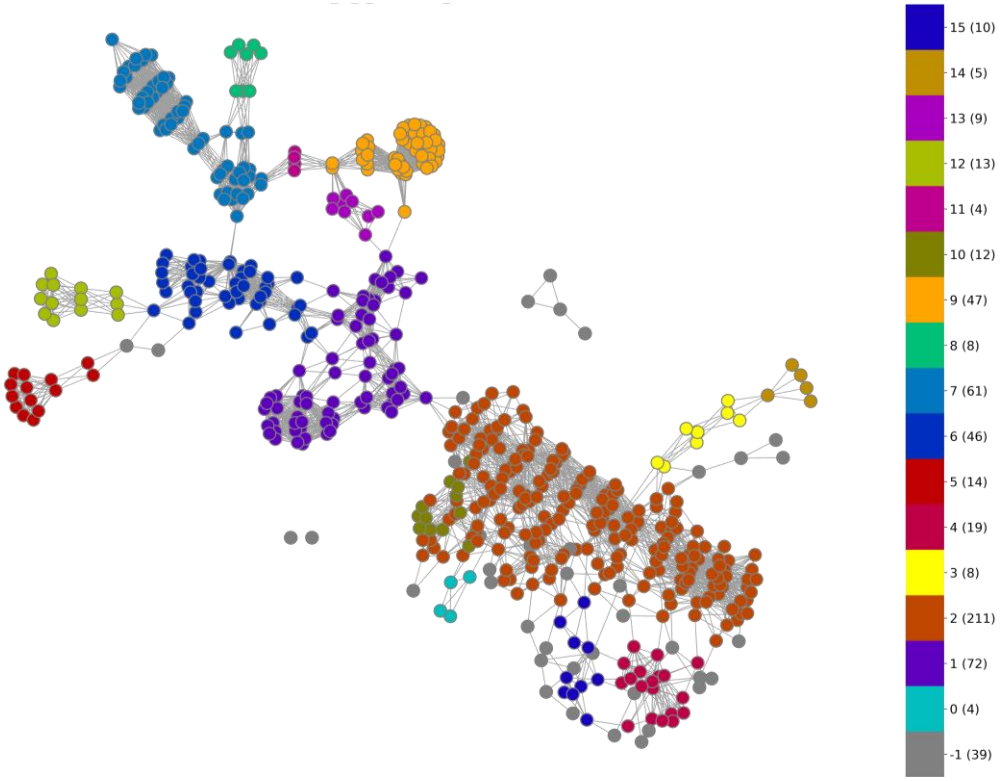


Figure 22. **The output of the 4-clique percolation method**

We also consider the mean community trends built by visit rates (**Figure 24**) to illustrate how the TDA graph stratifies the data. The trends are well aligned with the completion status coloring on **Figure 23 a)** and indicate how the percolation communities differ by the outcomes.

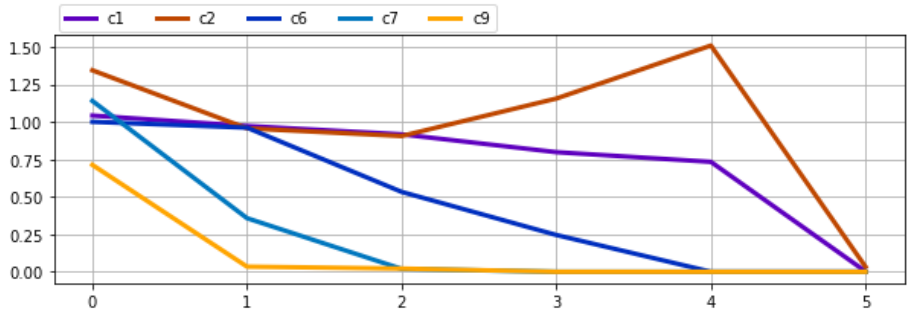


Figure 24. **Mean trends of five biggest communities by visit rates**

The subgroups of **Figure 25** can further be studied by utilizing standard statistical methods to determine the predictors that may be responsible for the similarity of the attendance pattern observed within the identified subgroup of patients.

Let us incorporate predictors in order to give examples of pattern explanation found by the clique percolation method.

It is discovered that the **sixth** community significantly differs by a gender predictor from all the rest nodes of the graph as well as by the pairwise comparison (see **Figure 26**). While men and women are almost equally distributed in the **sixth** community, the majority of the participants in the study were men. It is also established that even more men are in the **first** community (significant difference, p-value < 0.05, with the rest of the dataset).

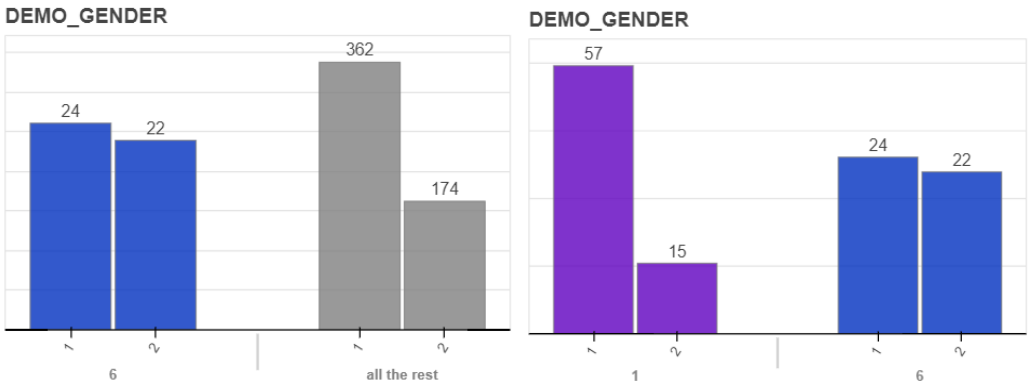


Figure 26. Differences in a gender

This **first** community has also differences in a socio-living distribution predictor (with a significantly bigger percent of those who live in controlled environment) as compared to the **fourth** community as well as it differs by a socio-occupation predictor from the **second** (see **Figure 27**), **seventh** and **ninth** communities.

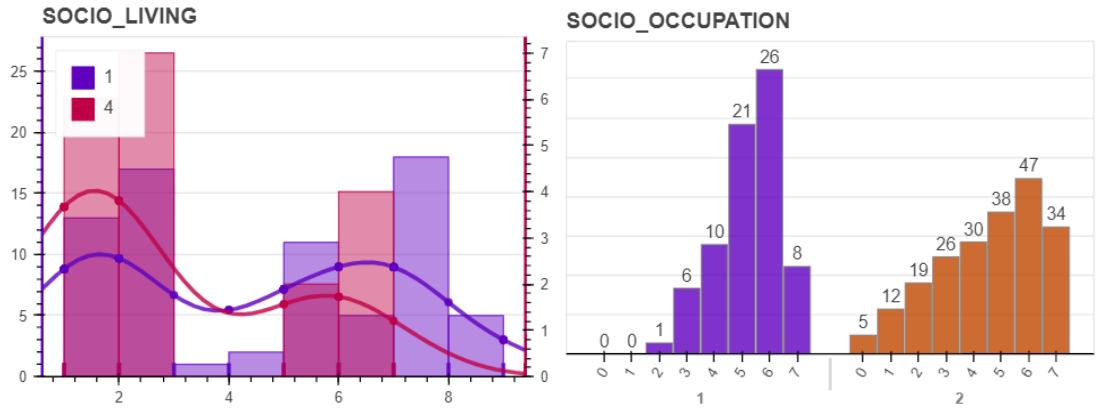


Figure 27. Differences in socio-living and socio-occupation distributions for the first community

The **first**, **second** and **forth** communities significantly differ by a socio predictor with the majority having an automobile compared to the **seventh**, **ninth** and **fifteenth** communities (**Figure 28**), where most of the people does not have a car.

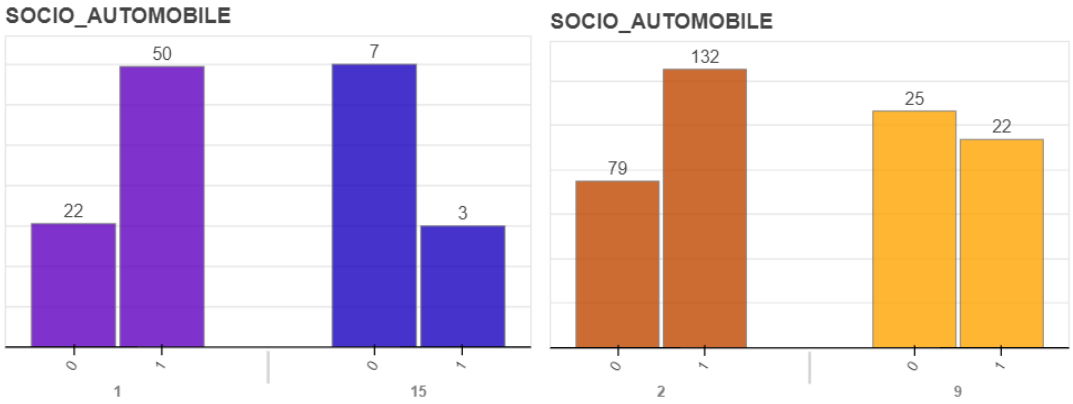


Figure 28. Differences in automobile available

Another example of the significant differences of the **ninth** community is by a demographical predictor of religion in comparison to all the rest of the dataset as well as in the pairwise comparison, with both the **first** and **second** communities (**Figure 29**).

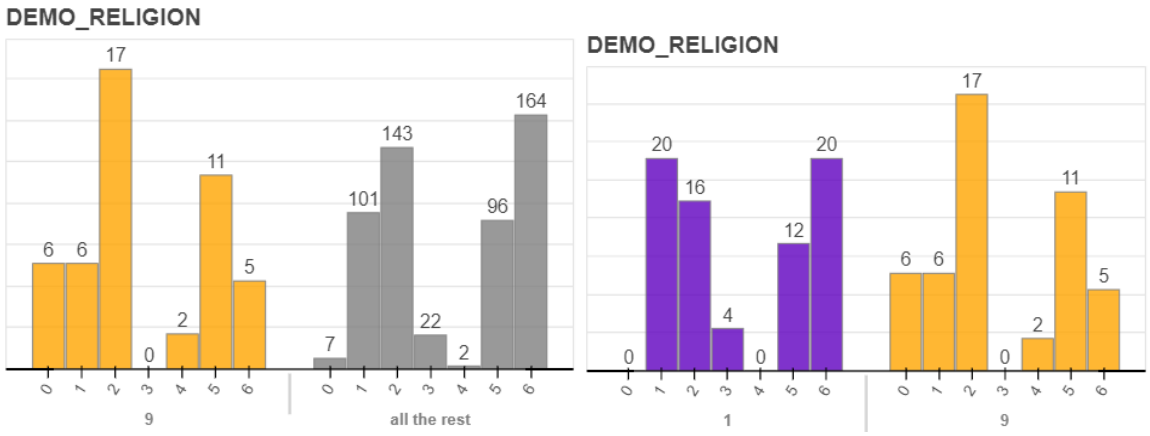


Figure 29. Differences in religion

The presented analysis with applying visual discovery tools can be implemented on any other topological model built on predefined outcomes of interest by various predictors.

4. BIAS DETECTION WITH TDA

The experiment concerns the bias detection in data collected during a clinical study at a center. A “center” is also referred herein to as a “site” and means a place of collection of data. For this research, we select some variables, which are measured for each patient during some visits, then we contaminate the data separately by variables and by centers and, finally, compare the ability of common statistical methods as well as TDA to detect the fact of contamination.

For the experiment, we selected 339 patients out of 582 who completed at least 10 visits (including the baseline visit) during the first 9 weeks of the study. Among the selected 339 patients, there were those who had one missed visit during these weeks. Data on this missed visit were completed using the Last Observation Carried Forward (LOCF) method. The variables measured during these 9 visits included both binary variables (the presence or the absence of drugs in the urine, the presence or the absence of certain symptoms of opiate addiction, etc.) and continuous variables (temperature, pulse, blood pressure (BP), respiratory rate, weight, etc.). Data on the presence of drugs, as well as data on the presence of symptoms of opiate dependence were aggregated in such a way as to obtain new indicators, namely the number of different drugs in the urine and the number of different symptoms of opiate dependence by visit. Continuous variables were left unchanged. The total number of variables is 9.

4.1. STATISTICAL METHODS FOR BIAS DETECTION

The problem of bias and fraud detection in data of multicenter clinical trials is considered from a statistical point of view in many sources (see [10], [11], [12])). The main idea is to use statistical criteria to identify “atypical behavior” of the center. To do this, a number of variables measured in patients are considered, and for each of these variables a comparison is made of the patients of the center versus all other trial patients. The null hypothesis is that the variable distribution among the center patients should be the same as the distribution across the entire population. A high level of significance ($p < 0.05$), especially if observed for a particular center across multiple variables, may indicate that the center's data are biased, contaminated, inaccurately collected, or fabricated. As it is said in [12], a very large number of possible statistical methods challenges the selection of the most appropriate methods to identify unusual patterns at centers. So we concentrated on the most common tests, like the t-test, F-test for variances, and Mann-Whitney-Wilcoxon test for continuous variables, which provided almost the same results (further we consider only results of the Mann-Whitney-Wilcoxon test). For discrete variables, the χ^2-test was applied.

Also note that in typical situations, statistical methods are univariate. However, in our experiment, the data consist of repeated measurements of variables. Therefore, for the statistical comparison of the center versus all other centers, we had to aggregate the measurement results of each variable within a patient. For continuous variables, we calculated the mean for each patient, while for discrete variables, we calculated the 0.8 quantile for each patient and discretized it again by dividing the obtained quantiles into several intervals.

4.2. TOPOLOGICAL MODELS FOR BIAS DETECTION

Bias detection in data using TDA is based on the hypothesis that the nodes of the graph corresponding to patients from the affected center tend to group, see **Figure 30**.

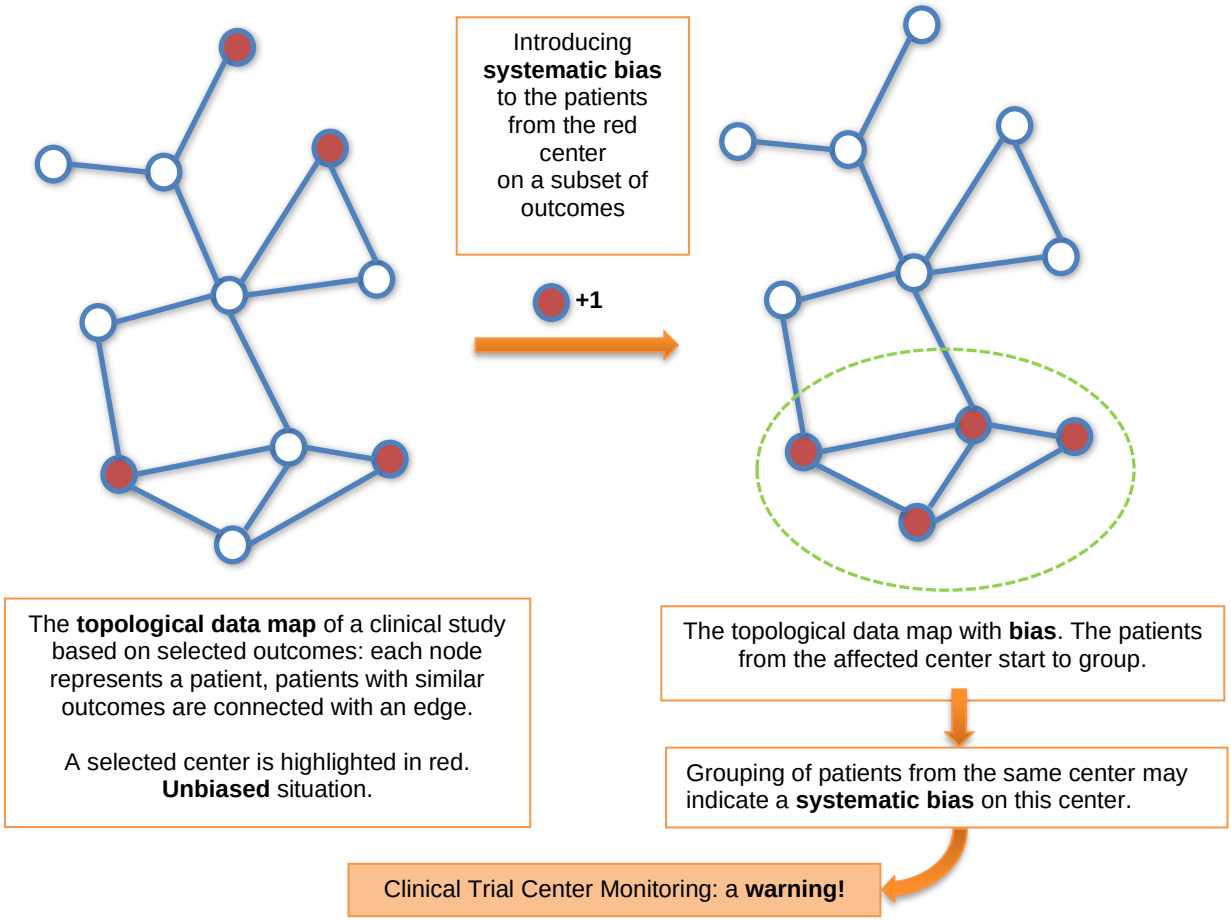


Figure 30. Bias detection in clinical trial central monitoring

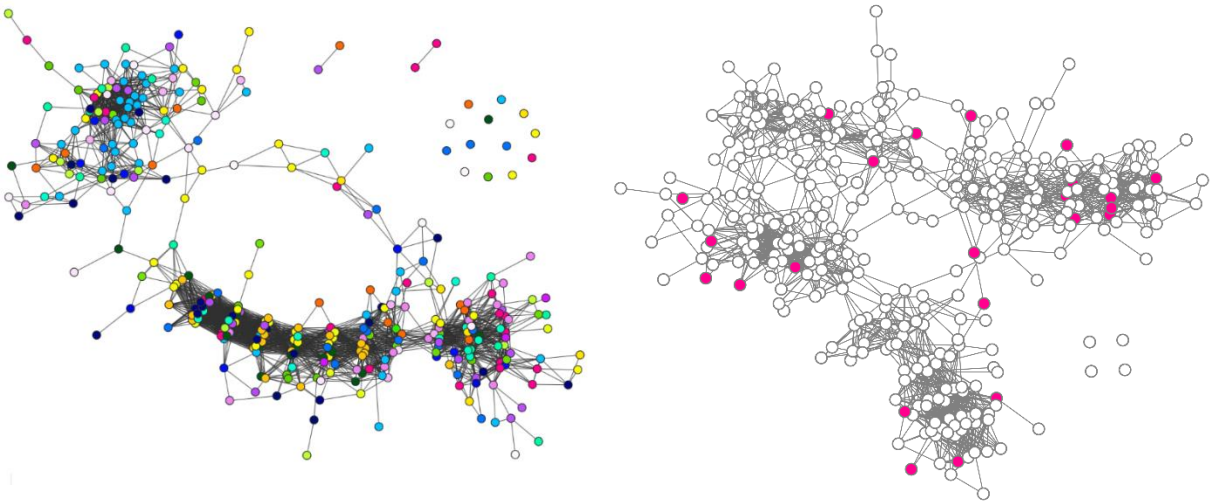
The TDA approach potentially offers several advantages compared to statistical methods, namely:

- TDA allows for visually identifying data grouping without manipulation of numbers and tables.
- TDA enables visualization of even small subsets of patients, for which statistical methods are not suitable.
- TDA enables building the model by using any type of variables also for combination of variables of different types, statistical methods should be carefully selected with respect to the types of variables of interest.
- TDA allows considering multiple of variables simultaneously when constructing a topological model (e.g., results of measurements of the same variable over multiple visits), whereas commonly used statistical methods are univariate. As noted in [10] , "multivariate statistical techniques offer more checking possibilities, but they are seldom used in clinical trials, if at all."

The topological model of the data is constructed separately for each of the selected variables. The vector, representing each patient, is given as

=	Baseline	Visit 1	Visit 2	Visit 3	…	Visit 9

Each of the 9 variable-based graphs includes 339 patient nodes connected by edges if they have similar variable dynamics. The Manhattan and Euclidean metrics and eccentricity and density projections were used in constructing the graphs. For each variable-based graph, we highlighted each of the 10 major centers to obtain 90 center-colored variable-based graphs, see **Figure 31**.



a) by number of opiate symptoms, nodes coloring by centers;

b) center-colored graph by diastolic BP, center 771 is highlighted

Figure 31. Variable-based graphs

4.3. EXPERIMENT RESULTS AND DISCUSSION

In the experiment, we contaminated the data by each of the 10 major centers according to the following schemes:

For continuous variables:

- Shift by a constant value (2 variants)
- Shift by a value following a normal distribution (2 variants)
- Transformation of data given by $\exp(\ln(x + 0.5) + 2sd) - 0.5$, where sd is the standard deviation of log-transformed values of x. This transformation was considered in [11] in testing statistical methods for fraud detection.

For discrete variables:

- Shift by a constant value
- Replacing all the values within the patient with their median.

Thus, we obtained 10 centers × 5 schemes × 6 variables = 300 contaminated datasets for continuous variables and 10 centers × 2 schemes × 3 variables = 60 contaminated datasets for discrete variables.

It is worth noting that introducing bias in many cases cannot be easily detected through simple observation, as biased variable values often fall within the normal ranges of these variables.

For all contaminated datasets, new variable-based graphs were constructed. The assessment of the degree of grouping of nodes within biased center was conducted by two methods. We evaluated the degree of grouping subjectively by visually dividing the graphs into three categories: with no grouping, with mild grouping, and with strong grouping of the center nodes, as well as using two objective node grouping indices: the *connectivity index*, which represents the fraction of edges connecting patients of the same center to the maximum possible number of edges connecting these patients, and the *average clustering coefficient* of center nodes, where the node clustering coefficient is the fraction of possible triangles through that node that exist.

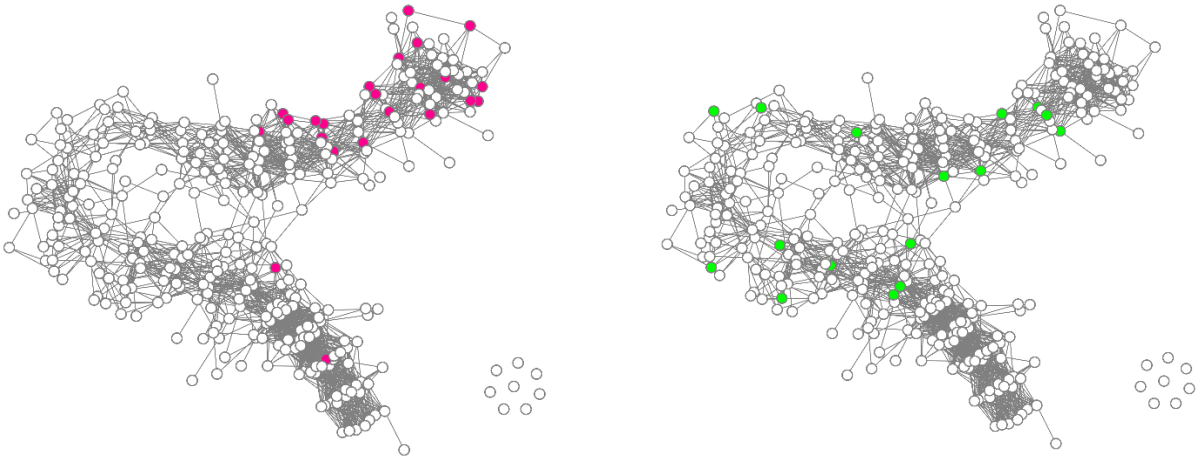
From the statistical point of view, the main difference between unbiased and biased cases should be in absence and presence of significant difference of a center from others ($p > 0.05$ and $p < 0.05$, respectively). From the TDA point of view, the biased case should demonstrate stronger grouping of

center nodes than the unbiased one. The main results of the experiment depending on p-value combinations of unbiased/biased cases are summarized in Table 1.

		Unbiased variable p-value / Biased variable p-value							
		Case I		Case II		Case III		Case IV	
		$> 0.05 / < 0.05$		$> 0.05 / > 0.05$		$< 0.05 / < 0.05$		$< 0.05 / > 0.05$	
			%		%		%		%
Number of cases (out of 360)		137	38.1	74	20.5	121	33.6	28	7.8
of them	Demonstrate increasing node grouping indices	124	90.5	50	67.6	104	86.0	8	28.6
	Can be detected visually	99	72.2	16	21.6	71	58.7	2	7.1

Table 1. TDA vis statistical methods in bias detection

A statistically significant difference of a center from other centers ($p < 0.05$) can be seen in the graph by the grouping (mild or strong) of nodes belonging to that center. The absence of node grouping is well-aligned with the inability to reject the null hypothesis ($p > 0.05$). In 78 out of 90 (86.7%) unbiased center-colored variable-based graphs, grouping of center nodes corresponds to significant statistical difference of the center or ungrouping corresponds to insignificance, see **Figure 32**.

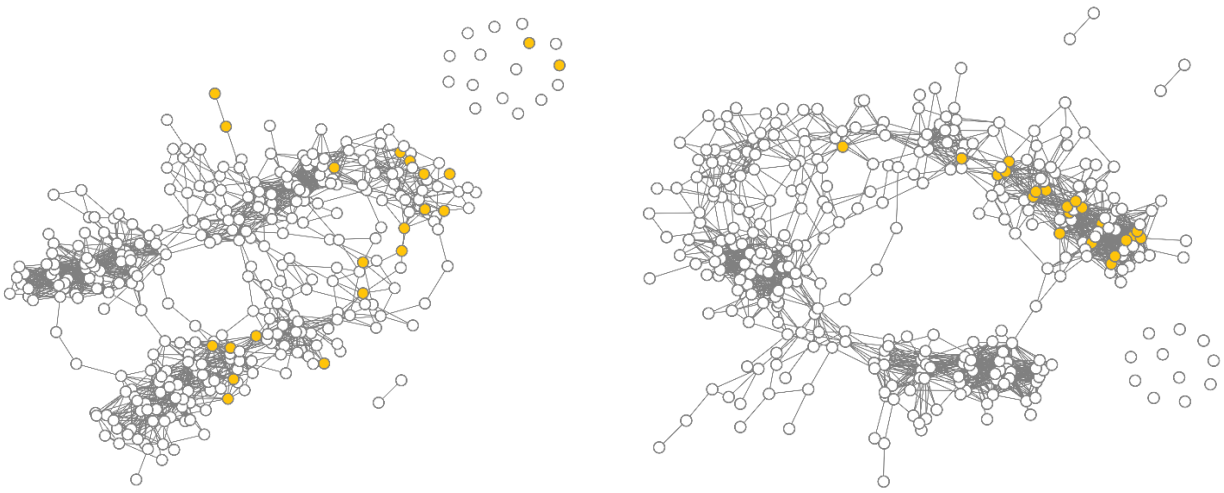


a) systolic BP, center 771; $p < 0.01$; center nodes mild grouping

b) systolic BP, center 539, $p = 0.344$; no center nodes grouping

Figure 32. Correspondence of significant difference and center nodes grouping on graphs

188 cases out of 360 biased cases (52%) are visually detected by graphs (grouping of the center nodes in the biased case is stronger than in the unbiased one) and in 286 cases (79%) graph clustering indices increase in the biased case, as shown in **Figure 33**.



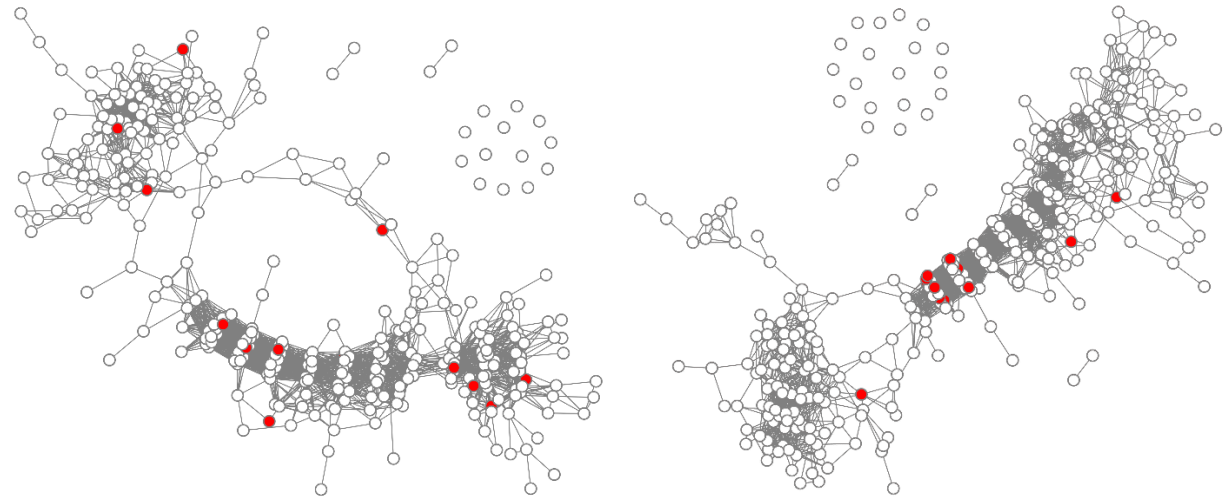
a) unbiased data, connectivity index=0.03, $p = 0.177$

b) biased data, connectivity index=0.28, $p < 0.01$

Figure 33. The variable-based graph by a pulse rate. The patients from center 622 are colored. The bias introduced is a constant shift by +10.

258 cases out of 360 (72%) demonstrate significant difference of biased centers (at a p-value of 0.05), although, only in 137 cases (38.1%) centers behavior before and after bias introduction is natural ($p > 0.05$ before contamination and $p < 0.05$ after, see Case I in Table 1). In 127 cases (33.6%) statistical approach does not differentiate the unbiased and biased cases (both p-values are < 0.05), but in 71 (58.7%) cases out of these 121 the biased variable-based graph demonstrates stronger center nodes grouping (see Case III in Table 1).

There are some cases in which the statistical approach does not allow to draw a conclusion about bias (in biased and unbiased cases both p-values are > 0.05, see Case II in Table 1), but TDA visualization demonstrates stronger center nodes grouping in the biased case, see **Figure 34**.

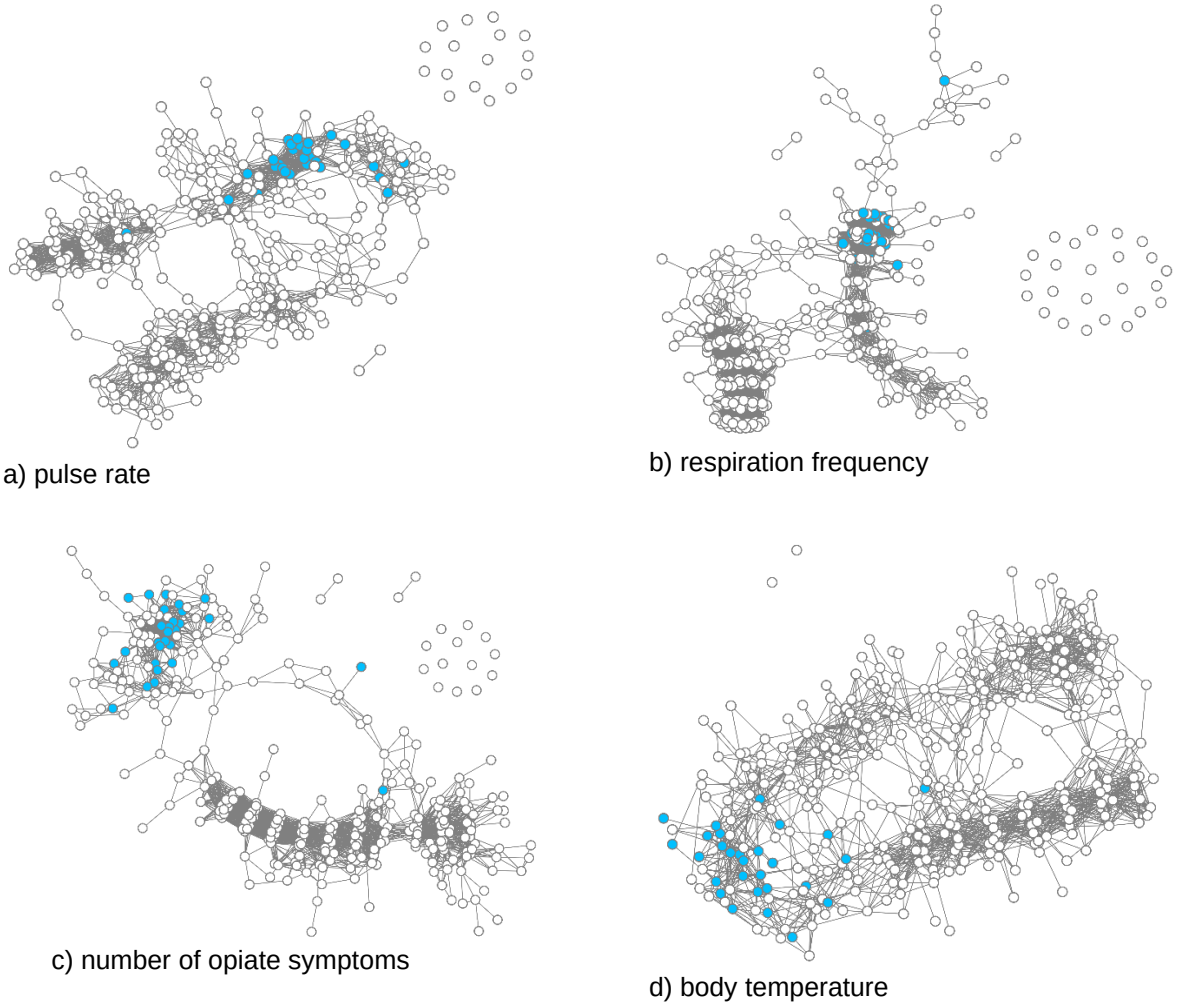


a) unbiased, number of opiate symptoms, center 862; $p = 0.77$; no grouping of center nodes

b) biased, number of opiate symptoms, center 862; $p = 0.13$; mild grouping of center nodes

Figure 34. Center nodes grouping in the biased case, although both unbiased and biased p-values are > 0.05

It is worth noting that many centers statistically significant differ from the others ($p < 0.05$) even in unbiased data. It may be caused by data particularities, patient selection (339 out of 582), data preprocessing, or other reasons. At the same time, we discovered that one of the centers (center 178) significantly differs from the others and demonstrates strong center nodes grouping in 7 out of 9 variables, which is a signal to check the data gathering and processing at this center, as shown in **Figure 35**



a) pulse rate

b) respiration frequency

c) number of opiate symptoms

d) body temperature

Figure 35. The variable-based graph with unbiased data for center 178. The similarity of patients from center 178 by a number of variables is a signal to check the data integrity in this center

Thus we conclude that TDA effectively detects bias in cases where statistical methods can also identify it, and it can also discover the data contamination in some cases where statistical methods do not show differences between biased and unbiased cases (both p-values are < 0.05 or both p-values are > 0.05). Thus, an advantage of the TDA approach is its visual component.

CONCLUSION

In this paper, a novel graphical method for visual discovery in Risk Based Monitoring was discussed. The method, based on Topological Data Analysis (TDA), produces a topological model of a dataset in the form of a graph. In this graph, a node represents a single patient, while two nodes are connected with an edge in the corresponding patients are close with respect to some variables of interest. Several features can be incorporated into the graph, and the model can present multivariable data in one structure, thus allowing the researcher to focus on various aspects of the study at once. By visually analyzing the graph,

and further using some ML algorithms, various parts in the graph can identified as interesting. These parts in the graph could be further analyzed with statistical methods.

Risk Based Monitoring is an important component of modern clinical trials. The elements of RBM are not fully standardized and it presents a significant challenge for clinical researchers and stakeholders.

In this paper, we apply the graphical method of TDA to address several challenges within RBM, namely, patience retention (as a part of Key Risk Indicators) and remote monitoring of activity in clinical centers. Two experiments were carried out. In the first, a topological model for patience retention and drop-out rates was analyzed. In the second, a procedure for detection of systematic bias (artificially introduced or already present) using visual discovery was discussed. The method was further compared to the classical statistical methods. It was demonstrated that in many cases, due to intrinsically multivariate nature of the topological model, the TDA approach has an advantage over the classical statistical methods which are often focus only on one single variable.

REFERENCES

[1] FDA, "Oversight of Clinical Investigations — A Risk-Based Approach to Monitoring. Guidance for Industry.," 2013. [Online]. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/oversight-clinical-investigations-risk-based-approach-monitoring. [Accessed 30 November 2023].

[2] B. Barnes, N. Stansbury, D. Brown and oth., "Risk-Based Monitoring in Clinical Trials: Past, Present, and Future," *Therapeutic Innovation & Regulatory Science,* vol. 55, p. 899–906, 2021.

[3] "National Institute on Drug Abuse; NIDA-CSP-1018," [Online]. Available: https://datashare.nida.nih.gov/study/nida-csp-1018. [Accessed 30 November 2023].

[4] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society,* vol. 46, no. 2, pp. 255-308, 2009.

[5] H. Edelsbrunner and J. Harer, Computational Topology: An Introduction, American Mathematical Society, 2010.

[6] S. Evans, "Statistical aspects of the detection of fraud," in *Fraud and Misconduct in Biomedical Research (3rd Edition)*, W. F. Lock S., Ed., London, BMJ Books, 2001.

[7] Borg, I.; Groenen, P., Modern Multidimensional Scaling: theory and applications (2nd ed.), New York: Springer-Verlag, 2005, pp. 207-212.

[8] L. v. d. Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research,* pp. 2579-2605, Nov 2008.

[9] G. Wu, S. Childress, Z. Wang, M. Roumaya, C. Stern, C. Dickens and J. Wildfire, "Good Statistical Monitoring: A Flexible Open-Source Tool to Detect Risks in Clinical Trials," *Therapeutic Innovation & Regulatory Science,* Posted 15 Jan, 2024.

[10] M. Buyse, S. Evans and oth., "The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials," *Statistics in Medicine,* vol. 18, no. 24, p. 3435–3451, 1999.

[11] L. Trotta, Y. Kabeya, M. Buyse and oth., "Detection of atypical data in multicenter clinical trials using unsupervised statistical monitoring," *Clin Trials.,* vol. 16, no. 5, pp. 512-522, 2019.

[12] J. M. Pogue, D. P. J. K. Thorlund and S. Yusuf, "Central statistical monitoring: detecting fraud in clinical trials," *Clin Trials,* vol. 10, no. 2, pp. 225-35, 2013.

[13] A. Zomorodian, Topology for Computing, Cambridge: Cambridge University Press, 2005.

ACKNOWLEDGMENTS

We would like to acknowledge **Oleksandr Leonov** (Kharkiv National University, Ukraine), **Lyudmyla Polyakova** (Kharkiv National University, Ukraine), and **Victoriia Shevtsova** (Intego Group, Ukraine) for

being core members of the research team and handling the computational experiment. Without you, this research and the experiment would not have been possible.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Contact: Sergey Glushakov
Company: Intego Group
Address: 2300 Maitland Center Pkwy, Suite 240, Maitland, FL 32751, USA
Work Phone: +1 407 512-1006
Email: sergey.glushakov@intego-group.com
Web: www.intego-group.com