PhUSE 2017

Paper TT10

Topology-based Clinical Data Mining: Identifying Hidden Patterns in Clinical Datasets

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

ABSTRACT

The emerging field of clinical data mining seeks to apply data integration and visualization tools that can improve our understanding of the entire clinical dataset. In this paper, we show how a novel methodology – topology-based clinical data mining (TCDM) – can be used to reveal multivariate patterns in clinical trial datasets. TCDM allows investigators to extract comprehensive topological maps of the data without first having to develop a model or hypothesis. This geometric, data-driven approach allows researchers to identify meaningful relationships in data that would otherwise be left unidentified by traditional biostatistical approaches. The TCDM methodology was adopted to develop a prototype of a software platform that provides a computational environment in which researchers can perform data mining experiments on clinical datasets. We discuss the key steps of the TCDM workflow and describe the surprising results of computational experiments that were performed on several datasets.

INTRODUCTION

Topology-based Clinical Data Mining is a research project that began over a year ago as a result of a partnership between Intego Group, a Functional Service Provider, and Kharkiv National University, one of the top Universities in Ukraine. This presentation discusses the key steps involved in the TCDM workflow: data integration, generation of topological data maps, visual inspection of interesting data maps, statistical analysis, and interpretation of discovered relationships.

Performing comprehensive analysis of a clinical trial dataset can be challenging. While the majority of approaches to mining clinical data focus on univariate relationships between a specific outcome and a few predictive variables, there is a lack of data integration and visualization tools available that can improve our understanding of the entire dataset. Examining clinical data with a focus on a specific outcome in isolation from other factors may lead to an incomplete, or even misleading, view of complex settings. Standard biostatistical methods are used as technical tools to confirm (or refute) the hypotheses generated by an investigator and, hence, rely on the researcher's ability to develop solid hypotheses. However, in the case of clinical trial datasets, the number of possible hypotheses to explore is very large, and it can be very difficult to select the most valuable.

In this paper, we describe the application of a novel holistic approach – topology-based clinical data mining (TCDM) – to discover multivariate patterns in clinical trial outcomes. TCDM allows an investigator to extract comprehensive topological maps of the data without first having to develop a model or hypothesis. A topological map provides a compressed, graphical representation of a multidimensional set of interrelated clinical outcomes. It zooms in on robust, geometric properties of the data that do not change under small perturbations or deformations. The robust data patterns and relationships, which remain invariant under (properly defined) small perturbations of the data, are referred to as topological properties. This focus on topological properties is what makes TCDM less sensitive to noise and, thus, helps to identify and visualize the key features of a clinical trial dataset despite the risk of errors, irregularities or missing values occurring in the 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 establishments 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. In Section 2, the mathematical foundations of the topological approach are briefly introduced. First, we explain what information can be encoded in a topological data map and how to construct the latter. We then discuss methods of pre-processing an original clinical trial dataset obtained in CDISC SDTM or ADaM format as a means of preparing a table of clinical outcomes – a synthetic dataset used as an input to the TCDM workflow. Next, the general TCDM workflow is outlined. The section concludes with a short description of a software prototype that provides a computational environment in which researchers can perform data mining experiments on clinical datasets. Section 3 describes a case study in which TCDM was applied to a real world clinical trial (the treatment of cocaine dependence). We examine the implementation of the full cycle of TCDM methodology in practice, from raw data preprocessing to the interpretation of the findings. We conclude by presenting a brief comparison of TCDM with the traditional biostatistical methodology.

2. METHODS

2.1. TOPOLOGY AND DATA MINING

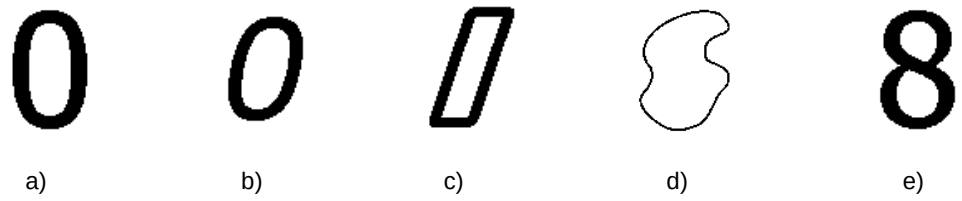
Topology is a field of mathematics that deals with the properties of objects that remain invariant under continuous deformation. Imagine a surface that is made of very thin and elastic material. You can bend, stretch or crumple the surface any way you like; however, you can not tear it or glue any parts of it together. As you deform the surface, it will change in many ways, but some properties will remain the same. The idea that underpins topology is that some geometric properties depend not on the exact shape of an object, but rather on how its parts 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 zero are equivalent since they can all be transformed into each other in a continuous way without cutting or gluing (Figure 1 a-d). It is possible to change the size, thickness, or slope of the digit 0 by a continuous deformation; however, one property remains invariant: The object separates the plane into two regions, interior, and exterior. At the same time, 0 is not topologically equivalent to 1 or 8: 1 does not encircle a region, and 8 contains two holes (Figure 1e). The topological classification of the digits 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 that are taken from distinct classes are equivalent from the topological point of view.

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 one removes any two points from it, it will become 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.



a) b) c) d) e)

Figure 1. Different representations of the digit 0 (a-d) are topologically equivalent.
All of them share a common topological property: They divide the plane into an interior region and an exterior region. The digit 8 (e) is not equivalent to 0 since it encloses two internal regions.

2.1.1. HOW CAN TOPOLOGY FACILITATE THE UNDERSTANDING OF COMPLEX DATA?

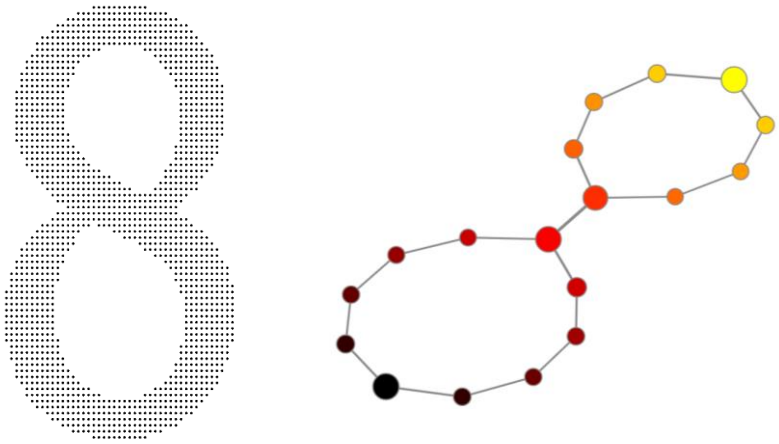
Topology deals with abstract mathematical entities, such as curves and surfaces, that consist of an infinite number of points. In practice, however, all datasets are necessarily finite. Recently, a new field has emerged at the crossroad of topology and data science. Topological data analysis (TDA) aims 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 a multidimensional space) at multiple scales or resolutions, from fine- to coarse-grained. Given a complex dataset, TDA is used to extrapolate the underlying topology and build a compressed, yet comprehensive, topological summary of the dataset. TDA exploits a variety of methods and algorithms stemming from computational topology and geometry, statistics, and data mining. For detailed expositions of the mathematical theories that underpin TDA together with some applications in biology see [5-7] 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 indistinguishable locally; however, they are not equivalent if they are considered as a whole. Topology offers a toolbox by which local information about an object can be integrated into a global summary. Thus, topology can provide a 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 is always noisy.

2.1.2. WHAT IS A TOPOLOGICAL DATA MAP?

In Figure 2a, the digit 8 is shown as a granular cloud, consisting of hundreds of dots. Every dot corresponds to a pixel and can be uniquely determined by its coordinates (x_1, x_2) on a two-dimensional grid. Suppose we are only interested in the fundamental qualitative properties of this geometric dataset and do not care about finer details, such as its exact shape or the variation in the intensity of the pixels. Using topological methods, we can build a compressed representation of the dataset in the form of a topological data map (Figure 2b). It contains only 18 nodes and 19 edges and captures the essential feature of the original figure; namely, that the latter consists of two loops.



a) b)

Figure 2. The digit 8 as a "granular dataset" (a) and its topological map (b)

A topological data map is a graphical representation of a dataset in which each node denotes a specific subgroup of data points. There is an edge between two nodes if, and only if, the respective subgroups of data points share common elements. A topological data map retains the relevant information about the dataset in a compact and efficient manner. One well-known class of topological maps is that of subway maps, which preserve the order of subway stops on each line and the interconnections between different lines.

In the context of clinical research, the dataset under study is typically a table of outcomes in a particular clinical trial. The table rows correspond to the individual participants in the clinical trial, and columns contain information on specific outcome measures of interest such as lab tests, vitals, questionnaires, etc. Given a table of clinical outcomes, two types of parameters are required to generate its topological data map. The first of these is a projection, a function that is used to stratify patients into subpopulations. The second is a distance function that measures the proximity between patients. The distance function makes it possible to split each subpopulation into clusters of related patients with similar outcomes.

To be considered for further analysis, a topological data map should meet certain requirements. Namely, it should:

- accurately represent the original dataset;
- eliminate the features of the dataset that are not relevant to the purpose of the study;
- reduce the complexity of the features that are shown on the data map;
- be insensitive to small noise such as errors of measurement.

When a suitable projection, distance, and scale of a topological data map are selected, some information will be eventually lost from the table of outcomes. For example, individual patients will be combined into clusters. The goal of TCDM is to create a data map that highlights the most significant and meaningful features of the original dataset while secondary unimportant features are eliminated from the constructed visualization.

2.2. PREDICTORS AND OUTCOMES IN TCDM

A very common situation in statistics occurs when the distribution of an outcome (or response variable) is related to one or several predictors (or explanatory variables). A standard approach through which researchers study the relationship between the predictor and the outcome is the application of a suitable statistical model. The model selection depends on the data types of the predictor and outcome (quantitative, binary, categorical, etc.) and often involves additional assumptions concerning the distribution of the outcome. For example, linear regression is often used when both the predictor and the outcome are quantitative (e.g., BMI and blood pressure); Fisher's exact test or χ^2 test can be applied when both variables are binary or categorical (e.g., gender and ECOG score); and logistic regression can be a suitable model for evaluating the relationship between a quantitative predictor and a binary outcome. The application of such approaches can be problematic in the context of complex settings that have multivariate outcomes; i.e., when many related outcomes are recorded for the same individuals.

TCDM is naturally designed to assist researchers to deal with multivariate heterogeneous outcomes in such a manner that it is possible to study several related outcomes of different types (quantitative, ordinal, categorical) together. An incomplete list of multivariate outcomes includes a series of repeated evaluations of some response variable over time; simultaneous evaluations of different, but potentially correlated biomarkers (e.g., levels of serum creatinine, blood urea nitrogen, and neutrophil gelatinase-associated lipocalin as a means of evaluating kidney function); and questionnaire data to assess patient's general health or quality of life, etc.

TCDM takes a panel of personalized outcomes of a clinical trial as its input. More specifically, the outcomes table is a synthetic dataset that consists of row vectors $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$, with each vector corresponding to a single participant. Here, x_i denotes the i-th outcome reading for the participant labeled $\boldsymbol{x}$. Outcomes are either calculated or directly extracted from original "raw" datasets that were collected during the course of the clinical trial and are stored in the CDISC SDTM or ADaM format.

From the clinical research perspective, an outcome is an evaluation of some aspect of a participant's health that results in a recorded datum. There is more than one way of classifying clinical trial outcomes (see Table 1). Depending on the research goal, it is useful to differentiate between outcomes linked to biomarkers and clinical outcome assessments (COA) (see [8]). A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention [9]. A COA is any assessment that may be influenced by human choices, judgment, or motivation, and it may provide either direct or indirect evidence of the benefits associated with a given treatment. In contrast to biomarkers, which are determined using automated processes or algorithms, COAs depend on a participant's or clinician's implementation, interpretation, and reporting of the data.

TABLE 1. CLASSIFICATIONS OF CLINICAL TRIAL OUTCOMES

Clinical Trial Goal	Specialty	CDISC Domain	Data Type	Variable Type
Safety Efficacy Effectiveness Quality of life	Allergy/Immunology Cardiology Endocrinology Gastroenterology Hematology/Oncology …	AE EG LB QS VS …	Cross-sectional Longitudinal Aggregate	Quantitative Categorical Ordinal Interval

It is important to note that specific research objectives require customized configurations of outcomes panels. In this paper, we consider several different outcomes panels derived from the same clinical trial dataset to study various aspects of the disease.

2.3. DISTANCES

Distance functions (or simply, distances) is a tool that is used in TCDM to measure the similarity between the data points that belong to a given outcomes dataset. A dataset endowed with a distance becomes a metric space that can be studied using geometric and topological methods. In mathematics, a distance on a set $\boldsymbol{X}$ is a function $d(\boldsymbol{x}, \boldsymbol{y})$ that quantifies proximity between each pair $(\boldsymbol{x}, \boldsymbol{y})$ of elements of $\boldsymbol{X}$. The most popular and intuitive example is, of course, the Euclidean distance:

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}$$

This measures the length of the straight-line segment between points $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ in an n-dimensional Euclidean space. In many applications, the underlying set, $\boldsymbol{X}$, does not have an Euclidean structure or, even it if does, the Euclidean distance is not the most suitable for the problem at hand. Below, we give several examples of distances that can be applicable for the configurations of outcomes tables that typically arise in TCDM.

Normalized Euclidean Distance

The normalized Euclidean distance is represented by:

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\left(\frac{x_1 - y_1}{s_1}\right)^2 + \left(\frac{x_2 - y_2}{s_2}\right)^2 + \ldots + \left(\frac{x_n - y_n}{s_n}\right)^2}$$

where s_i is a scaling parameter, which is usually the standard deviation of x_i and y_i over the sample set.

The normalized distance helps to avoid the effect of units in which different outcomes are measured. Normalization should not be applied when all outcomes are expressed in the same units, since doing so may dramatically reduce large effects (when an outcome with a large contribution is divided by a large s_i).

Manhattan Distance

The Manhattan distance is represented by:

$$d(\boldsymbol{x},\boldsymbol{y}) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n|$$

It is more robust than the Euclidean distance because it attributes less weight to an outlying value of any particular outcome. It is also better suited to high-dimensional data.

Hamming Distance

Hamming distance is one of many proposed distances for purely categorical data. It is defined as follows:

$$d(\boldsymbol{x},\boldsymbol{y}) = \frac{n - m}{n}$$

Here, m is the number of matches between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$; i.e., the number of outcomes, i, such that $x_i = y_i$, and n is the total number of outcomes in the table.

The situation becomes more complex when several outcomes of different types (quantitative, binary, categorical) are combined in a single outcomes table. In the case of mixed data, none of the distances described above are directly applicable, and one needs to use a more general measure of distance, such as the Gower distance [10]. We will not employ the more complex constructs of distances in this paper.

2.4. PROJECTIONS

A projection function (or simply, a projection) is a numerical function defined on data points. In the TCDM framework, projections are used to extract relevant information from a table of outcomes and to summarize this information in the form of a topological data map. In some regards, TCDM projections are very similar to geographic map projections. Their general characteristics are as follows:

- All projections represent the original data in a compressed form and, therefore, distort the data;
- Different projections highlight different features of the data (at the expense of other features);
- There is no limit to the number of projections that can be applied to a given dataset.

TCDM projections can be roughly divided into two classes: domain-dependent and universal projections. Domain-dependent projections are defined in terms of the dataset under study. For example, in the table of outcomes that represents complete blood count readings at the end of a given study, the contents of a specific column (e.g., hemoglobin or white blood cells) can be used as a projection. In contrast, universal projections do not depend on the structure or composition of a dataset and can be derived from statistics, computational geometry, pattern recognition, signal analysis, etc.

While, in principle, an arbitrary mathematical function can be used in TCDM as a projection, in practice, some functions lead to representative and easily interpretable data maps more often than others. We present below a few basic examples of universal, geometric projections that have proven to be helpful in various settings. In what follows, $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_N}$ denote the n-dimensional vector rows corresponding to data points, $\boldsymbol{X}$ is the $N \times n$ matrix (table of outcomes) composed of these rows, $\boldsymbol{x}$ is an arbitrary data point, and $d(\boldsymbol{x}, \boldsymbol{y})$ is a distance function on the dataset.

$\boldsymbol{L_p}$-centrality estimator projection

$\boldsymbol{L_p}$-centrality estimator is defined by the following expression:

$$M_p(\boldsymbol{x}) = \left(\frac{1}{N} \sum_{i=1}^{N} d(\boldsymbol{x}, \boldsymbol{x_i})^p \right)^{1/p}$$

The parameter $p \geq 1$ controls the relative contributions of large terms in the sum. In the case $p = 2$, the estimator assumes the minimal value on the spatial average of vectors $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_N}$:

$$\arg\min_{\boldsymbol{x}} M_2(\boldsymbol{x}) = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x_i}$$

Hence, $M_2(\boldsymbol{x})$ measures the proximity of a data point to the spatial average or 'spatial center' of the dataset. In the case $p = 1$, $M_1(\boldsymbol{x})$ is simply an averaged sum of the distances from $\boldsymbol{x}$ to the data points. It is minimized by the so-called spatial median, which is also an important statistical estimator of location. The $\boldsymbol{L_p}$-centrality estimator coincides with the usual median in the one-dimensional setting (i.e., when data points are scalars), but, in the general case, its analytic solution is not known in closed form. However, the explicit central point is not needed to compute $M_p(\boldsymbol{x})$. This estimator provides an intuitive, intrinsic measure of centrality (higher values of $M_p(\boldsymbol{x})$ imply that $\boldsymbol{x}$ is located further away from the 'dataset center'). It can also be used for categorical and mixed data.

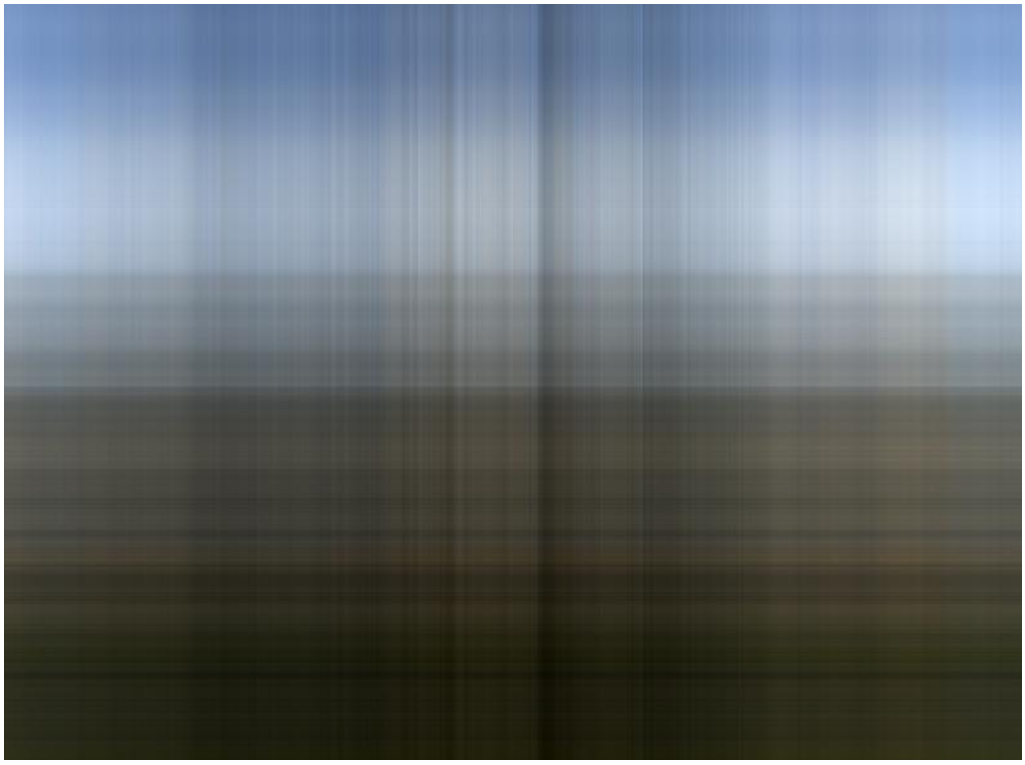
PC scores projections

Principal component analysis (PCA) is a statistical procedure that provides a series of best linear approximations for a dataset in which there are many correlated variables. PCA aims to reduce the dimensionality of the dataset while retaining as much of the variation that is present in the data as possible. For a thorough outline of the PCA theory and its various applications see [11].

Let's assume that the data matrix $\boldsymbol{X}$ is filled with purely quantitative values and the columns of $\boldsymbol{X}$ are centered (i.e., the sample mean of each column has been shifted to zero). We also assume that the data set is endowed with the usual Euclidean distance. By definition, the first PC score vector is the N-dimensional vector $\boldsymbol{X w_1}$, where $\boldsymbol{w_1}$ is the n-dimensional unit vector, which maximizes the sample variance of $\boldsymbol{Xw}$ in comparison to all other unit vectors:

$$\boldsymbol{w_1} = \arg \max_{d(\boldsymbol{w},\boldsymbol{w})=1} (\boldsymbol{Xw})^T \cdot (\boldsymbol{Xw})$$

a) The original image contains 800 x 593 pixels



b) Projection on the first principal component

Figure 3. Data compression by PCA.
Edinburgh Castle. Photograph by Kim Traynor, distributed under a CC BY-SA 3.0 license

$\boldsymbol{w_1}$ is referred to as the first loading vector. The coordinates of the vector, $\boldsymbol{Xw_1}$, are the first PC scores and they correspond to the respective data points. We define the first PC score projection on a data point $\boldsymbol{x}$ as the scalar product of $\boldsymbol{x}$ and $\boldsymbol{w_1}$:

$$t_1(\boldsymbol{x}) = \boldsymbol{x} \cdot \boldsymbol{w_1}^T$$

It can be shown that the first PC score vector inherits the maximum possible variance from $\boldsymbol{X}$. One can then progress to successively define the second, third, … n-th PC score vectors. In combination, PC score vectors provide best linear approximations to $\boldsymbol{X}$ of any rank. Figure 3 (a-b) shows the PC scores property to compress data and capture the maximum amount of variance on for a dataset representing a digital image of Edinburgh Castle.

The kernel density estimator projection

The kernel density estimator (KDE) is defined as follows:

$$\widehat{f_h}(\boldsymbol{x}) = \frac{1}{Nh^n} \sum_{i=1}^{N} K\left(\frac{d(\boldsymbol{x}, \boldsymbol{x_i})}{h}\right)$$

where h>0 is the bandwidth, and the kernel function, K, is a symmetric density. Usually K is the standard normal density function:

$$K(t) = (2\pi)^{-n/2} \exp\left(-\frac{t^2}{2}\right)$$

KDE is a generalization of the well-known histogram density estimator, where n-dimensional bins are replaced with smoothed 'bumps' centered at data points. The bandwidth, h, is a free smoothing parameter that determines the width of the 'bumps.' KDE is a statistical tool that can provide valuable information on the data distribution properties, such as skewness and multimodality (see [12]). Note that the multivariate KDE is not advised for small datasets in which the number of outcomes is greater than 5 (the required dataset size to ensure that the relative mean squared error is less that 0.1 is 768 for $n = 5$ and 2790 for $n = 6$). For further details, see Chapter 4 in [12].

2.5. TCDM FRAMEWORK

This section outlines the general TCDM workflow; i.e., how topological data maps of clinical trial datasets are generated and studied. The diagram presented in Figure 4 depicts the main stages of the TCDM. The process by which data maps are constructed consists of outcomes table preparation, selection of suitable distance and projection functions to reduce the outcomes table dimension, stratification of clinical trial participants into subgroups, partial clustering of participants within each subgroup, and topological features extraction based on the clustering results.

2.5.1. DRAWING A TOPOLOGICAL DATA MAP

To initiate the TCDM process, two synthetic datasets need to be determined: Outcomes and Predictors (see Section 2.2). In each dataset, a row represents a unique participant within the clinical trial, while the columns represent either observational variables (outcomes), such as safety and efficacy biomarkers, or predictors, such as demographic attributes, medical history, interventions, etc. Extracting the outcomes and predictors from the clinical trial data is a key part of the data pre-processing task. Depending on the objective of the TCDM research and the structure of the data, this task may also involve data cleaning and editing, transformation (e.g., conversion of categorical values into numerical data), normalization, and integration.

As described in Section 2.4, a projection function transforms data points (high-dimensional vector rows of outcomes table) to scalar numeric values. It is used to stratify participants into subgroups. The range of projection values is divided into several overlapping intervals. All individuals that are mapped by the projection to a specific interval define a subgroup of the total population of the clinical trial. The collection of subgroups constitutes a stratification grid of the dataset.

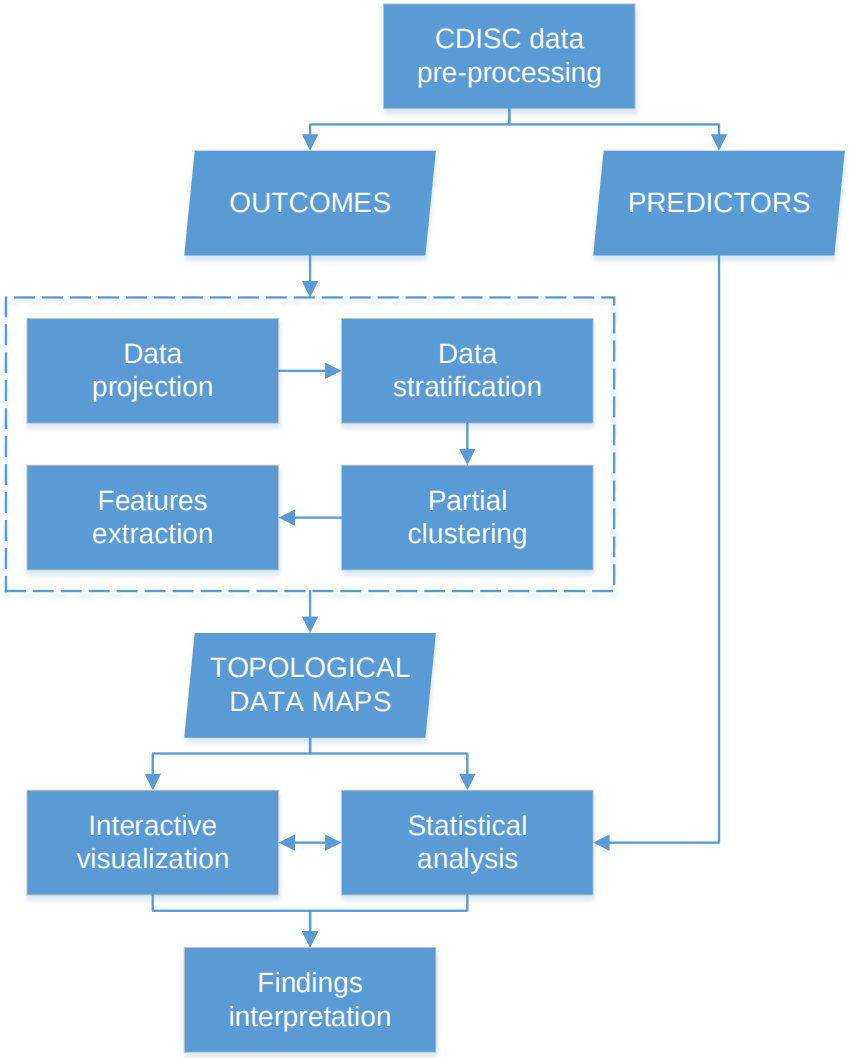
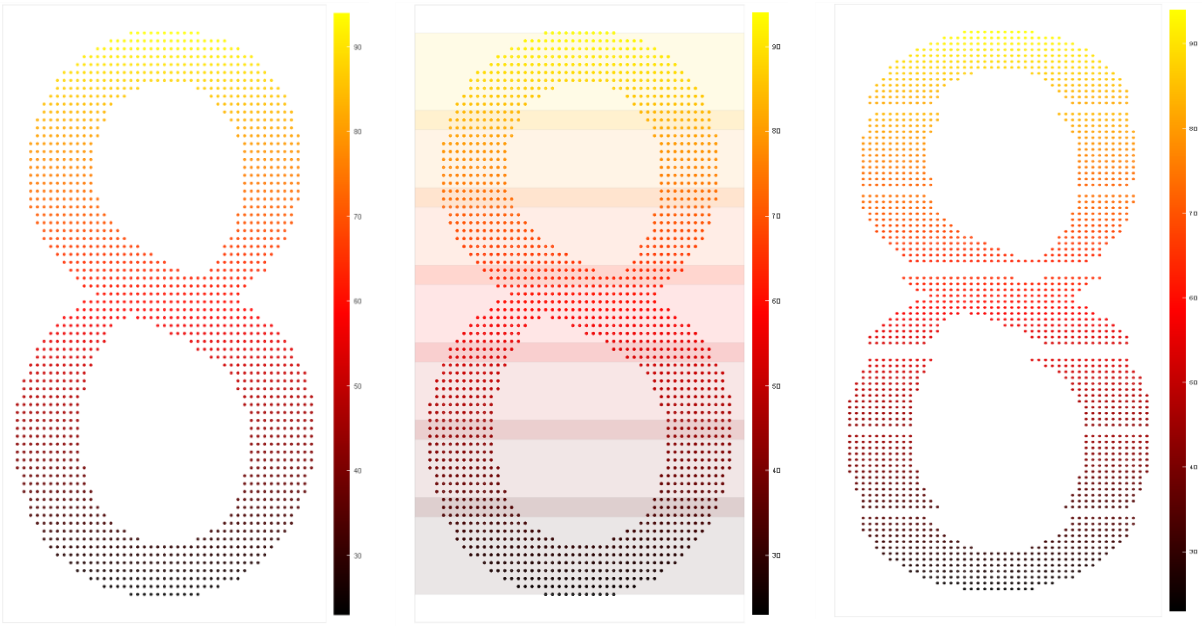


Figure 4. TCDM methodology workflow

The data stratification is controlled by several parameters, including the number of intervals, overlap of two adjacent intervals, balance (to quantify the degree of uniformity of the distribution of the data points across the stratification grid), etc. When one selects a relatively small number of intervals, one gets a coarse view of the data in which many individuals will be placed into a single subgroup within the stratification grid. A relatively large number of intervals lead to a fine-grained stratification, with just a few individuals in each subgroup. The number of intervals is determined according to the level of detail the researcher requires to be reflected in the topological representation of the dataset. Intuitively, the number of intervals controls the scale of a data map. A direct analogy between a data map and standard geographic maps can be drawn in that small-scale maps, such as those of the entire world or continents within it, show large areas of the Earth within a small space, while large-scale maps, such as city plans, show small areas in more detail.

After the data stratification is completed, each subgroup is clustered. Clustering is achieved by grouping individuals into even smaller subgroups (or clusters) such that the individuals in the same cluster have more features in common than those in other clusters. The similarity of individuals is determined by a distance function (see Section 2.3). The clustering task is solved by one of a whole family of algorithms that differ in terms of how the distances are computed. Apart from the choice of the distance function, one needs to specify the notion of what constitutes a cluster and how to identify such a cluster. The most appropriate clustering algorithm for a given dataset depends on the type of data (continuous, discrete, categorical, mixed, etc.) and often needs to be chosen experimentally.

During the final stage of data map construction, information about the clustering structure of all the subgroups that constitute the stratification grid is assembled and presented in the form of a graph. Since the dataset is stratified into overlapping subgroups, each data point can appear within several clusters. In the final graph, nodes correspond to clusters, and edges connect the nodes for which the respective clusters share common data points. Figure 5 illustrates how a data map is constructed for a simple two-dimensional geometric dataset.



a)

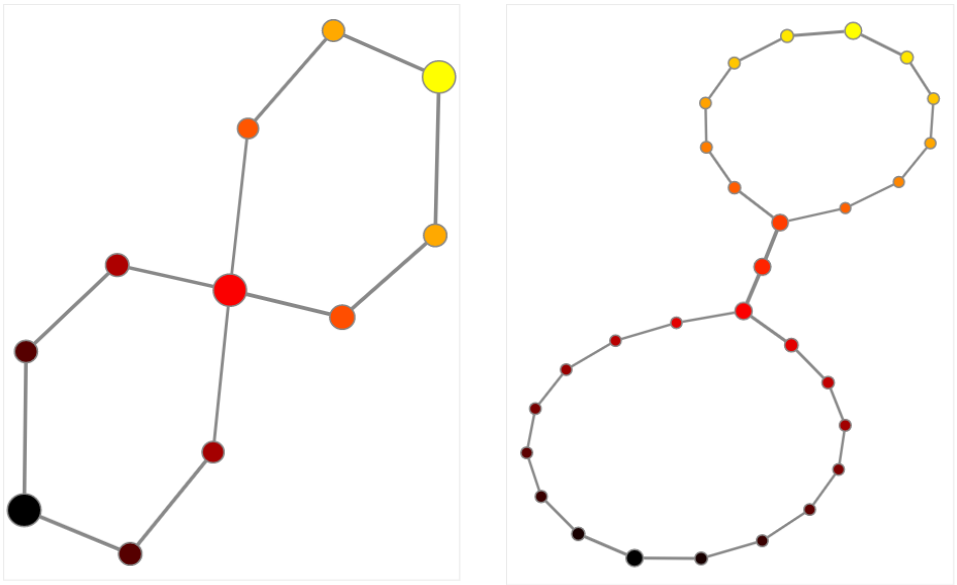
The original two-dimensional dataset on the plane. The projection maps a data point to its y-coordinate

b)

The stratification grid contains seven overlapping subgroups. The narrow pink horizontal bands contain points that are shared by two subgroups

c)

Clusters within subgroups



d)

A data map corresponding to the dataset stratification into 7 subgroups (a coarse-grained view of the dataset).

e)

A data map corresponding to the dataset stratification into 20 subgroups (a detailed view of the dataset).

Figure 5. Drawing a topological data map of a simple dataset

2.5.2. THE USE OF SOFTWARE FOR DATA MAP EXTRACTION AND ANALYSIS

The TCDM methodology was adopted to develop a prototype of a software platform that provides a computational environment in which researchers can perform data mining experiments on clinical datasets. The prototype implements the logic of the TCDM workflow (see Figure 4) and consists of a variety of scripts developed using Python, R, and SAS®. It relies on state-of-the-art machine learning algorithms, statistical tools, and data visualization libraries.

The transformation of 'raw' clinical trial data into predictors and outcomes datasets is performed by a dedicated SAS script. To launch the analysis, the researcher sets up the ranges of parameters required to construct the topological data maps that represent the outcomes table. The parameters include projection, distance, stratification grid, and others. Each combination of parameters produces a different data map. The resulting collection of data maps provides researchers with an opportunity to look at the dataset from different perspectives.

In the next step, the computing platform generates a large number of topological data maps (from a few thousands to several million) that correspond to the various combinations of parameters within the originally defined ranges. The 'features extraction' proprietary algorithm was developed to reduce the massive volume of topological data maps generated for a pre-defined set of parameters into a more narrow set of data maps that present meaningful insights. In other words, the features extraction module filters the most representative and stable data maps and reduces the collection size of the data maps from thousands, or even millions, to just a few dozen.

After a topological data map is extracted, the researcher visually explores the data map with the purpose of discovering interesting subgroups within the data. These subgroups can be further studied by utilizing standard statistical methods to determine the predictors that may be responsible for the similarity of responses observed within the identified subgroup of clinical trial participants.

Each node on a topological data map corresponds to a subgroup of clinical trial participants while, at the same time, nodes that share the same participants are connected. The size of the node illustrates the number of participants, and the width of the line connecting two adjacent nodes indicates the number of individuals in common. A node color depicts the median value of the projection for the group of individuals constituting the node.

The interactive visualization module provides researchers with an opportunity to manually perform a visual inspection of a topological data map that they can analyze to identify regions of interest. For example, the nodes that form fork-like structures or loops might be of interest for further research. In addition, isolated components or highly concentrated groups of nodes that form communities may indicate meaningful relationships in the outcomes dataset. While performing a visual inspection, the researcher can also re-color the data map in accordance with the median value of an outcome or predictor selected from the corresponding datasets generated at the beginning of the TCDM workflow. The use of color codes may highlight how a subgroup of individuals represented by a given region of the data map might be different from the rest of the clinical trial participants.

The researcher can select any region of the data map that exhibits interesting geometric properties to perform a further statistical analysis. After running statistical tests, a table of predictors with the corresponding p-values can be calculated to determine if the distribution of the predictors for the selected subgroup of participants is different to that of the rest of the study population. If the desired significance level of any predictor is found to be significant, the researcher can construct a histogram that represents normalized frequency distributions of the predictor for both the individuals in the selected region of the map and the rest of the population. At any time, interesting findings can be bookmarked and documented for further in-depth analysis by a clinical study group.

3. RESULTS

In this section, we demonstrate the applications of TCDM using data from a large-scale clinical study that was conducted with the support of the NIDA Clinical Trials Network (CTN-0048). The aim of the study was to investigate the safety and effectiveness of buprenorphine in the presence of naltrexone for the treatment of cocaine dependence (Study data and documents are available for download at https://datashare.nida.nih.gov/study/nida-ctn-0048). During this double-blind, placebo-controlled study, a sample population of 302 participants at 11 sites were randomly assigned to one of three conditions: Experimental Group 1: 4 mg buprenorphine plus naltrexone, Experimental Group 2: 16 mg buprenorphine plus naltrexone, and Control: placebo plus naltrexone. The eligible participants were males and females between 18 and 65 years of age who met DSM-IV criteria for cocaine dependence and exhibited past or current opioid dependence or abuse. Participants were treated with pharmacotherapy for eight weeks, with three clinic visits per week. Follow-up assessments occurred at one month and three months post-treatment.

The primary outcome was urine drug screen (UDS)-corrected, self-reported cocaine use during the last four weeks of treatment. Secondary outcomes assessed cocaine use by UDS, medication adherence, retention, and adverse events. Our analysis concentrated predominantly on the efficacy assessments of the study, which were represented by UDS and various psychometric measures such as score on the Visual Analog craving scale (VAS) for opioids and cocaine, the Beck Depression Inventory (BDI), treatment effect assessment (TEA), etc.

3.1. PREDICTORS AND STATISTICAL TESTS

A total of 18 univariate characteristics of the study participants were selected as the primary predictors for the quantitative evaluation of the statistical association between these characteristics and the study outcomes (Table 2). Of these, one variable contained a unique identifier of a participating site, three predictors corresponded to basic demographic data (patient's gender, ethnicity, and date of birth), five variables were binary flags of specific substance use in the past 12 months, and the remaining nine predictors characterized substance abuse and dependence according to the current DSM-IV diagnostic schema.

TABLE 2. PREDICTORS

Variable Name	Interpretation
SITE	Community treatment program site
DEBRTHDT	Date of birth
DEGENDER	Participant gender
DEETHNIC	Participant ethnicity
DSDAG	Age when first experienced at least three symptoms of substance dependence
DSAAG	Age when first experienced at least one symptom of substance abuse
DSUSALCH	Used alcohol
DSUSAMP	Used amphetamines
DSUSMJNA	Used marijuana
DSUSBENZ	Used benzodiazepines
DSUSOTHR	Used other substances
DSOFALCO	Often used alcohol
DSOFCOCA	Often used cocaine
DSOFAMPH	Often used amphetamines
DSOFMJNA	Often used marijuana
DSOFOPIA	Often used opiates
DSOFBENZ	Often used benzodiazepines
DSOFOTHR	Often used other substances

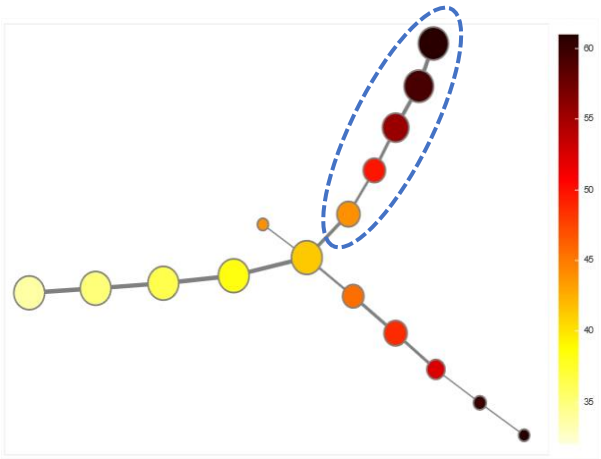
For the purpose of the statistical analyses, we distinguished between continuous, mixed, binary, and categorical (non-binary) univariate predictors according to the variable type. Continuous predictors were examined using the standard two-sample Kolmogorov-Smirnov test. This method verifies whether two data samples are obtained from the same distribution. The test assumes that the underlying distributions are continuous (no ties in the variable's values are allowed). In the event the distributions were not entirely continuous in the current study, we conducted a bootstrap version of the Kolmogorov-Smirnov test that has previously been found to yield accurate results in the more general setting (see [13]). To examine the statistical association between two samples within the categorical data, we used Fisher's exact test and χ^2 test for the binary and non-binary categorical variables respectively.

3.2. ANALYSIS OF TOPOLOGICAL DATA MAPS

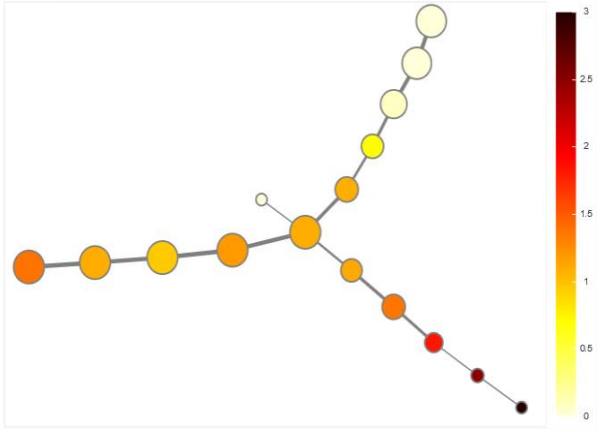
We analyzed several data maps obtained for various subsets of efficacy outcomes. In each case, the L_p-centrality estimator was employed as a projection (see Section 2.4). This projection allowed us to evaluate and visualize the stratification of patients according to their proximity to the 'spatial center' of the population. The 'spatial center' can be viewed as a generalized multivariate version of the population average or median.

URINE DRUG SCREEN (UDS)

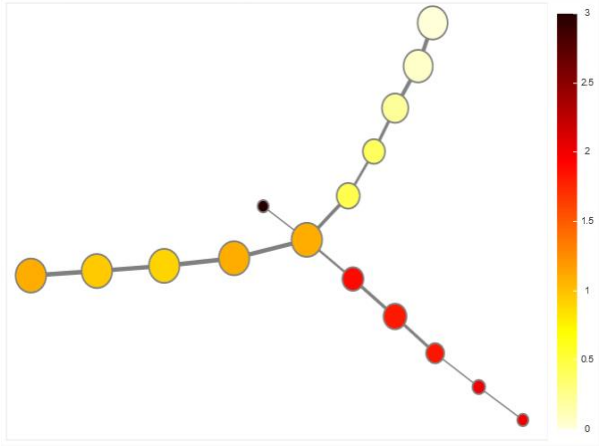
The UDS was conducted at every clinic visit to test for the presence of the following drugs: opiates, oxycodone, barbiturates, benzodiazepines, cocaine, amphetamines, methamphetamines, marijuana, methadone, and Ecstasy. For the purpose of this analysis, each outcome corresponded to the number of substances indicated by the UDS (the UDS score) during a given visit. This resulted in the development of a 27-dimensional dataset that consisted of vector rows of the UDS scores at various times (baseline, thrice-weekly assessments over the 8-week active medication period, the 1-month follow-up, and 3-month follow-up).



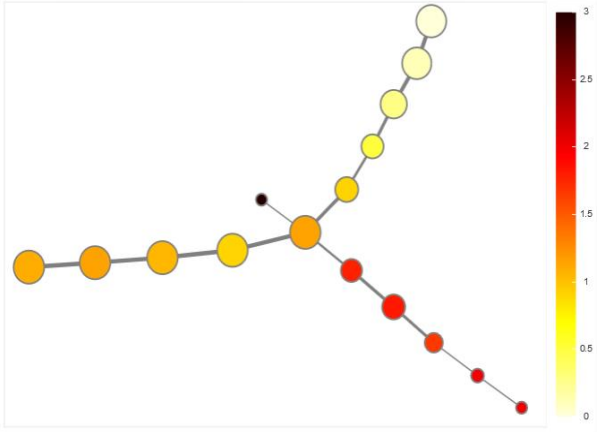
a)
The colors of the nodes correspond
to the median values of the L_p-centrality projection



b)
UDS score at baseline



c)
UDS score on visit 501

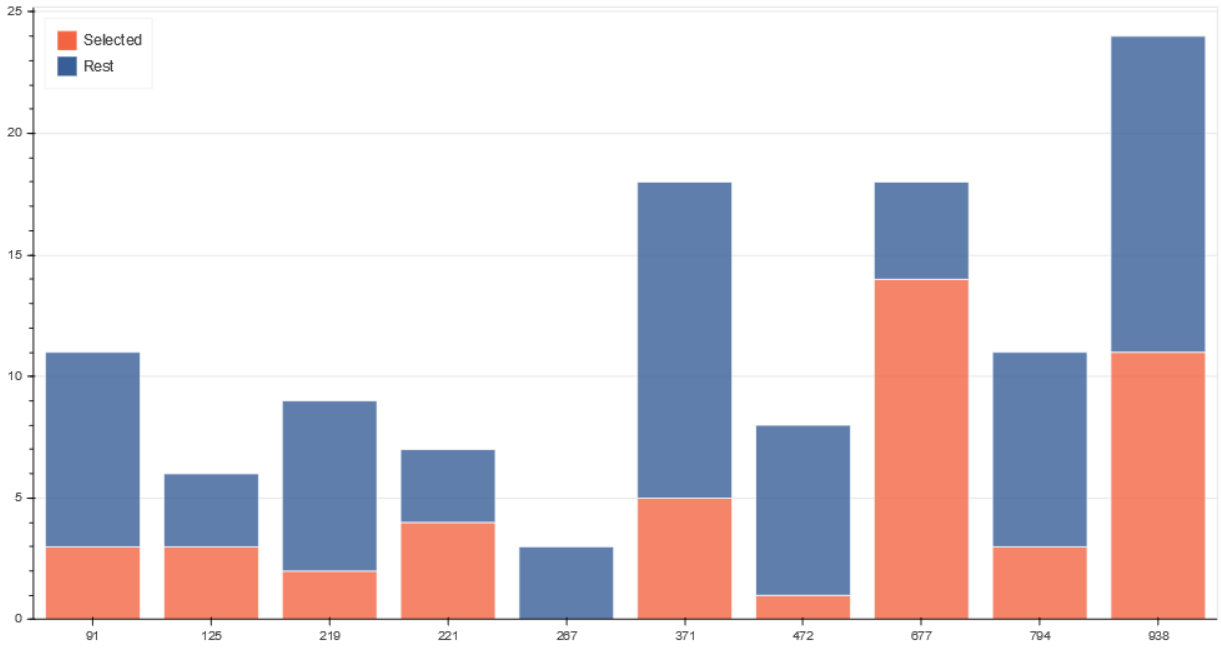


d)
UDS score at 3 months follow-up

Figure 6. Topological data map of the UDS score table

Predictor	Type	Pvalue
DSOFALCO	binary	0.00107
SITE	discrete	0.0153
DSUSBENZ	binary	0.0344
DSOFBENZ	binary	0.0363
DSOFMJNA	binary	0.0379
DEBRTHDT	mixed	0.097
DSUSMJNA	binary	0.157
DSOFOPIA	binary	0.255
DEETHNIC	binary	0.312
DEGENDER	binary	0.326

a)
Table of predictors and their p-values calculated to assess discrepancy between the selected group of patients constituting the low UDS score tine (46 patients) and the rest of the population (69 patients).



b)
Bar plot of the SITE variable for the selected group and the remaining patients.

Figure 7. Statistical analysis of predictors to account for discrepancies in the outcomes

Figure 6a represents a topological data map of the table of outcomes containing n =115 patients whose UDS records were available for all visits (the UDS was not collected for the remaining 187 patients at one or more visits). Note that the 'tines' of the fork-like part of the graph correspond to patients who exhibited higher projection values than those observed in the rest of the population. This implies that the patients in the tines were located further away from the population spatial center (in terms of the distance between the vectors of the UDS scores). We recolored the data map according to the value of the UDS score at baseline (UDSVIS00) and observed that one of the tines was predominantly populated by patients with a low number of substances detected in their urine (average UDSVIS00 $= 0.28$) while patients in the other tine had consistently higher numbers of substances detected in their urine (average UDSVIS00 $= 1.25$), see Figure 6b. This pattern persisted for UDS scores during other visits. As the time progressed, the average scores of patients with relatively high scores at baseline remained high, and the average scores of patients with relatively low scores at baseline remained low (Figure 6 c-d).

We grouped together all patients in the 'low UDS score' tine (46 patients) and compared them with the rest of the population using non-parametric statistical tests on the set of predictors defined in Section 3.1. The analysis revealed several statistically significant predictors ($P < 0.05$) that could account for the observed discrepancy in the UDS outcomes: DSOFALCO, SITE, DSUSBENZ, DSOFBENZ, DSOFMJNA (see Figure 7).

3.3. EXPLORATORY ANALYSIS USING METRIC GRAPHS

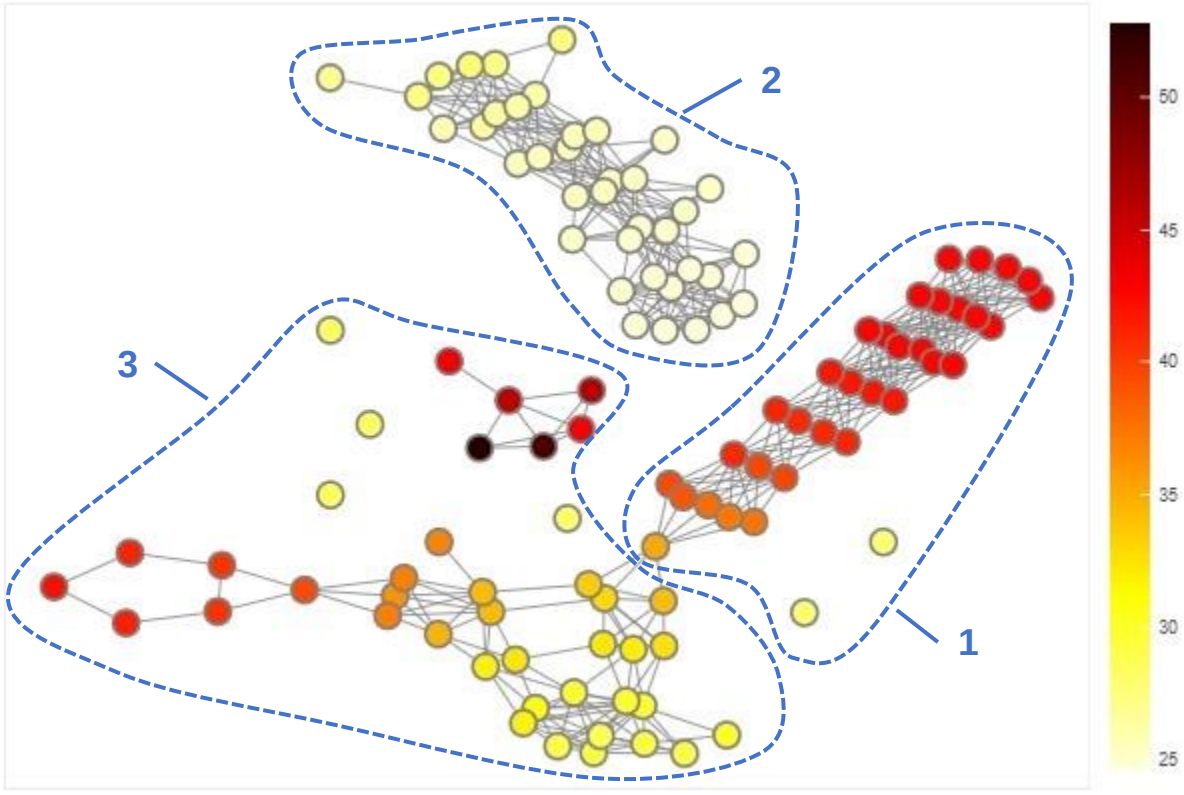
A further round of computational experiment was conducted using metric graphs, which represent topological data maps in which every node corresponds to one patient in the population. This gave us the ability to explore the metric graph constructed of the UDS scores (Figure 8). The same color coding was used in the metric graphs as that applied in the topological map employed in the previous rounds of the experiment. Based on the subjective visual exploration, it was assumed that there were three potential communities of patients and that these exhibited different patterns over time in terms of UDS (Figure 8a). Community 1 represented those patients who entered the study with a history of light drug abuse. Their score remained at a similar level throughout the study (Figure 8b). The response to therapy that the patients of this community exhibited appeared to be positive. The way in which the members of Community 2 and Community 3 responded to therapy was not as good as that of Community 1; however, the response of Community 2 was better than that of Community 3 (Figure 8c).

After identifying these three groups of patients based on visual exploration and different colorings of the metric graph, the further statistical analysis was performed in SAS, and the following unexpected results were identified:

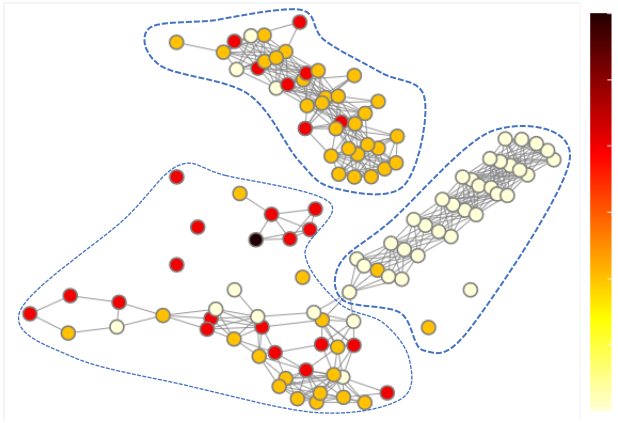
- Community 1: On average, the patients of this group were older than those of Community 2 and Community 3.
- Community 3: Community 3 contained more patients who had a history of benzodiazepines abuse in comparison to the patients in Community 2.
- Community 2: The patients of Community 2 exhibited a lower depression score throughout the study in comparison to those in Community 3. They also had a history of physical and neurotic disorders.

All these points are non-obvious from the exploratory point of view and are worth further investigation by the research team.

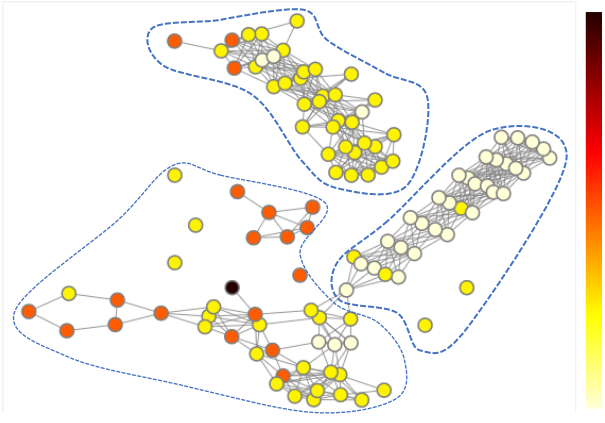
NOTE. More information about different rounds of experiment conducted by the research team can be found at PhUSE Wiki page related to this presentation.



a)
Three potential communities of patients



b)
The color of the nodes indicates
the baseline UDS score



c)
The color of the nodes indicates
the end of study UDS score

Figure 8. Metric graph represents a topological data map in which every node corresponds to one patient

CONCLUSION

In this paper, we presented a novel, topology-based methodology (TCDM) that allows researchers to gain visual insights into clinical trial data and generate new exploratory hypotheses using these insights. The approach was illustrated using a real-world clinical study in which the straightforward application of TCDM led to several interesting findings that would otherwise have been difficult to identify through the use of standard statistical methods alone.

Table 3 summarizes the fundamental features of the TCDM and the ways in which it is different to the majority of the alternative data-analytic approaches that are currently used in clinical research.

TABLE 3. COMPARISON OF TCDM AND STANDARD STATISTICAL METHODOLOGIES

Standard Statistical Approach	**TCDM**
Outcome variables are studied separately	Allows researchers to integrate and visualize multiple interrelated outcomes (see Figure 9)
Applicable under certain assumptions and/or requires a data model	Assumption-free and model-independent
Hypothesis-driven	Data-driven, hypothesis-generating
Focused on hypothesis testing for sub-groups and risk factors specified in advance by the investigator (e.g., stratum, gender, race, etc.)	Facilitates explicit identification of sub-groups of patients with similar outcomes

TCDM provides a flexible framework of data-driven, model-independent methods that can be readily adapted to various clinical research goals. It can be used both as an autonomous data-exploratory tool as well as an enhancement and extension of the traditional confirmatory analysis.

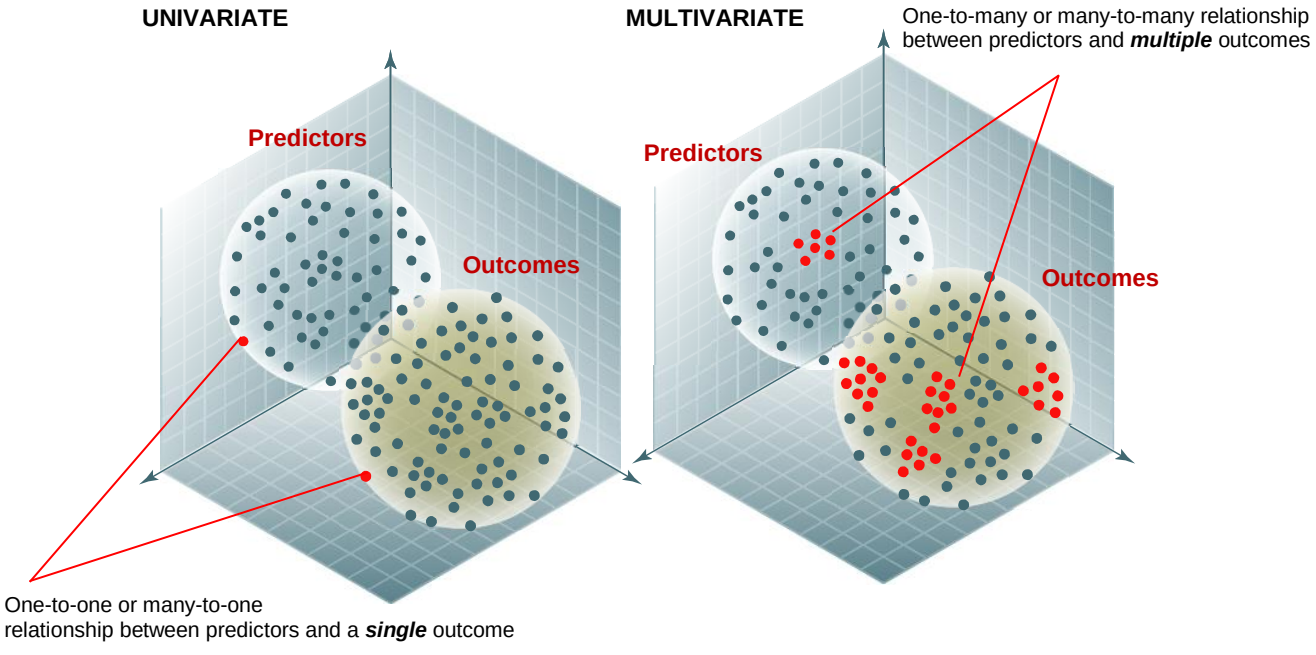


Figure 9. TCDM allows to visualize and study multiple clinical outcomes simultaneously

REFERENCES

[1] Nicolau, M., A. J. Levine, and G. Carlsson. 2011. “Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival.” Proceedings of the National Academy of Sciences 108.17 (2011): 7265-270. PNAS 2011 108 (17) 7265-7270.

[2] Nielson, J.L., Paquette, J., Liu, A.W., Guandique, C.F., Tovar, C.A., Inoue, T., et al. “Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury.” Nature Communications 6, 8581 (2015): n.pag. 10.1038/ncomms9581. pmid:26466022

[3] Torres, Brenda Y., Jose Henrique M. Oliveira, Ann Thomas Tate, Poonam Rath, Katherine Cumnock, and David S. Schneider. “Tracking Resilience to Infections by Mapping Disease Space.” PLOS Biology E1002494 14.4 (2016): n. pag. 10.1371/journal.pbio.1002494 .

[4] Alagappan, Muthuraman, Dadi Jiang, Nicholas Denko, and Albert C. Koong. “A Multimodal Data Analysis Approach for Targeted Drug Discovery Involving Topological Data Analysis (TDA).” Advances in Experimental Medicine and Biology Tumor Microenvironment 899 (2016): 253-68. 10.1007/978-3-319-26666-4_15. Springer.

[5] Edelsbrunner, Herbert; Harer, John (2010). Computational Topology: An Introduction. American Mathematical Soc. ISBN 9780821849255.

[6] Zomorodian, Afra (2005). Topology for Computing. Cambridge University Press. ISBN: 9780511546945.

[7] Carlsson, Gunnar (2009). "Topology and data". Bulletin of the American Mathematical Society. 46(2): 255-308.

[8] Qualification Process for Drug Development Tools (2014). U.S. Department of Health and Human Services. Food and Drug Administration Center for Drug Evaluation and Research (CDER). p. 1-35.

[9] Biomarkers Definitions Working Group (2001). Clinical Pharmacology and Therapeutics, 69, p. 89-95.

[10] Gower, J. C. (1971). A general coefficient of similarity and some of its properties. Biometrics, Vol. 27, No. 4. pp. 857-871.

[11] I.T. Joliffe (2002) Principal Component Analysis, 2nd ed. Springer Series in Statistics.

[12] B.W. Silverman (1986) Density Estimation for Statistics and Data Analysis, Chapman and Hall.

[13] Abadie, Alberto. (2002) “Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models.” Journal of the American Statistical Association, 97:457 pp. 284-292.

ACKNOWLEDGMENTS

We would like to acknowledge Bogdan Chornomaz (Kharkiv National University, Ukraine) 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.

We would also like to acknowledge Vladlen Kuzminov (Intego Group) for his involvement in the front-end development of the software user interface and the implementation of the topological data maps visualization module.

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