A Visual Analytics Framework for Analysis of Patient Trajectories

Kaniz Fatema Madhobi
Washington State University
Pullman, Washington
kanizfatema.madhobi@wsu.edu

Methun Kamruzzaman
Washington State University
Pullman, Washington
md.kamruzzaman@wsu.edu

Ananth Kalyanaraman
Washington State University
Pullman, Washington
ananth@wsu.edu

Eric Lofgren
Washington State University
Pullman, Washington
eric.lofgren@wsu.edu

Rebekah Moehring
Duke Center for Antimicrobial
Stewardship and Infection Prevention
rebekah.moehring@duke.edu

Bala Krishnamoorthy
Washington State University
Vancouver, Washington
kbala@wsu.edu

ABSTRACT

The problem of analyzing patient trajectories is fundamental to our ability to understand and characterize diseases and how we treat them in our hospitals, and to devise and explore effective alternative strategies for healthcare. In this paper, we present a new approach to analyze hospital patient trajectories. Based on visual analytics, our approach is aimed at aiding the domain scientist (in this case, a hospital bioinformatician or a data analyst) to visually navigate and analyze patient health trajectories in a scalable manner. More specifically, we view the problem as one of structure discovery and tracking how such structure evolves with time over the course of patients' stay at the hospital(s). An ability to scalably track and view the temporal progression of context variables associated with patients in conjunction with health indicator variables could provide vital clues on how practices affect outcomes. Furthermore, by enabling compact and consolidated views of complex patient trajectories, our approach can help to delineate subpopulations (i.e., subgroups of patients) that show divergent behavior. As a concrete case study in application and evaluation, we present results and initial findings on a large patient data set obtained from the Duke Antimicrobial Stewardship Outreach Network (DASON) database, with an aim of extracting factors relevant to antibiotic usage and stewardship in hospitals.

CCS CONCEPTS

• **Mathematics of computing → Paths and connectivity problems**; • **Information systems → Data mining**; • **Human-centered computing → Visualization**; • **Applied computing → Health informatics**.

KEYWORDS

Electronic Health Records, TDA, Patient Trajectories

ACM BCB '2019, Sept 7–10, 2019, Niagara Falls, NY

ACM Reference Format:
Kaniz Fatema Madhobi, Methun Kamruzzaman, Ananth Kalyanaraman, Eric Lofgren, Rebekah Moehring, and Bala Krishnamoorthy. 2019. A Visual Analytics Framework for Analysis of Patient Trajectories. In *Proceedings of ACM BCB '19: The 10th ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB '2019).* ACM, New York, NY, USA, 10 pages.

1 INTRODUCTION

The digitization of patient records has become a key instrument of change in the way biomedical healthcare is administered. Digitization has resulted in an abundance of data, and that has in turn resulted in an increased emphasis on scalable analytics and decision support systems that are primarily data-driven. Consequently, "data" in the form of patient electronic health records (EHRs) have exploded over the past decade [6]. While there are still a number of issues and challenges pertaining to the collection, formatting, curation, and integration of EHR data, from an analytical standpoint, one of the lead challenges in the area has been to generate analytical and computationally scalable frameworks for gleaning useful "information" from such data, and in the process aid and enable healthcare providers to improve the quality of decision-making.

In this paper, we focus on a specific type of electronic health record, namely patient trajectories, obtained from in-patient hospital records that typically cover a patient's stay at a hospital from the day of admission to the day of discharge. These data sets cover a wide array of treatment activities and all related meta-data associated with the health of a patient, as administered by caregivers as a function of time. Consequently, these data sets represent a treasure trove of information relating to understanding how a patient's health changes with every passing day at the hospital. For instance, these datasets can be very useful in the study of hospital-acquired infections (HAIs) [13, 15], or for analyzing conditions such as sepsis [12].

However, mining such information with actionable insights from hospital records can be significantly challenging owing to a number of factors, including but not limited to: size, variety, high dimensionality, ontology, etc. [5]. First, the *size* of these patient records in large hospital networks could be significantly large, covering possibly millions of patients treated across hundreds of hospitals and healthcare locations. Secondly, these large data sets also cover

a wide *variety* of patient conditions, treated in hospitals with different healthcare specialties and healthcare practices, and often different/unstandardized ways to gather patient data. Noise and missing data introduce an additional layer of complexity into the analysis of such data. Under these circumstances, trying to understand how treatment and healthcare practices affect patient outcomes and to devise effective strategies to help improve those outcomes, become challenging tasks. The tools and approaches that are currently used in the area are mostly database-oriented, where hospital informaticians store and retrieve data using hand-scripted queries and supplement them with custom pipelines that use standard statistical and regression tools for analysis.

Contributions: In this paper, we present an alternative approach to analyze hospital patient trajectories. Our approach, which is mathematically rooted in topological data analysis [17], is a visual analytics-based approach aimed at aiding the domain scientist (in this case, a hospital bioinformatician or a data analyst) to visually navigate and analyze patient health trajectories in a scalable manner. More specifically, we view the problem as one of "structure discovery" and tracking how such structure evolves with time over the course of patients' stay at the hospital(s). For instance, a patient could undergo different procedures, get administrated with a variety of drugs, change units within the hospital—all over the course of the stay; as the healthcare providers try to continually monitor and assess health risks and vulnerabilities. An ability to scalably track and view such temporal progression of context variables associated with patients, in conjunction with health indicator variables could provide vital clues on how practices affect outcomes—a key piece of information toward decision making at the coarser level of hospitals or units within hospitals. Furthermore, by enabling compact and consolidated views of complex patient trajectories, our approach can help in delineating subpopulations (i.e., subgroups of patients) that show divergent behavior.

As a concrete case study in application and evaluation, we present results and initial findings on a large patient data set obtained from the Duke Antimicrobial Stewardship Outreach Network (DASON) database [2] with an aim of extracting factors relevant to antibiotic usage and stewardship in hospitals. The DASON database contains a large collection of patient records from a network of 25 community hospitals curated by the Duke University School of Medicine. Although explained in this context, our approach is generalizable to analyzing patient trajectory data sets in other contexts.

The rest of the paper is organized as follows: Section 2 presents a brief overview of related works on patient trajectories and on topological data analysis applied to health analytics, along with a statement of how the framework presented in this work is different. In Section 3, we present our approach to problem modeling and describe our visual analytics framework. In Section 4, we present our results and findings on the DASON data set.

2 RELATED WORK

There have been many studies conducted on health registry, in particular relating to analyzing patient health records. However, the fraction of studies that focus on studying temporal trajectories have been relatively small. Giannoula et al. [4] and Jensen et al. [7] mined temporal patterns from patient disease histories using a time-analysis framework. Giannoula et al. [4] presented a method to identify common disease trajectories from electronic health records and, based on that information, cluster similar trajectories together. The core purpose is to find statistically significant disease associations in patients. Jensen et al. [7] presented a network representation of diseases to understand disease progression and to predict the probable next stage in a patient's life line.

The use of topological data analysis (TDA) in healthcare is relatively new. Utilizing topological concepts to find shape and pattern in data has made it an impactful tool to analyze health registry as well. Nicolau et al. [16] used TDA to analyze breast cancer transcriptional data. They identified a unique subgroup of patients with 100% survival rate. Li et al. [10] generated a patient-patient network from electronic health records, where each patient is a node and there is an edge between two nodes if they exhibit significant similar behavior (e.g., similar lab tests etc.). The authors used topological analysis to build this network, and identified three subtypes of Type 2 diabetes (T2D). They also analyzed disease comorbidities associated with each T2D subtype.

The work presented in this paper complements the above efforts. More specifically, we present a visual analytic framework that could be used to analyze and interact with large patient trajectory data sets acquired from hospitals. The results of applying our tool can help reveal, in an unsupervised manner, hidden higher-order structures about how different subpopulations within a large population show varied behavior, and how different factors possibly contribute to the variant behavior. This new analytical capability can provide valuable structural and behavioral insights into data that current pipelines are ill-equipped to reveal, and in the process could help us formulate better hypotheses from patient data.

3 APPROACH

In this section, we first present our approach for modeling the problem of analyzing hospital patient trajectories, identifying the different variables of interest, and the goals of analysis. Subsequently, we present our visual analytics framework for this problem. We present all our discussion viewing antimicrobial stewardship as our target application, as this application is used as a case-study throughout our study. However, the methodologies associated with the problem modeling as well as our visual analytics framework are both generalizable to other application contexts that involve patient trajectories.

3.1 Problem Modeling and Formulation

The goal of our antimicrobial stewardship study is to identify potential factors that contribute to antimicrobial exposure of patients in hospitals. We consider only in-patient data, i.e., for patients who are admitted and stay in the hospital for at least one day. The factors we consider can be broadly categorized into three classes:

- *Temporal:* length of stay (LOS), which is the number of days starting from the day of admission to the day of discharge or mortality for a given patient;
- *Spatial:* the hospital where the patient is admitted, and the hospital units where the patient receives care; and

- *Treatment-based:* agents and Standardized Antimicrobial Administration Ratio (SAAR) groups that a patient is exposed to over the course of their hospital stay.

The main performance (outcome) variable that we are interested in is *cumulative Days of Therapy (cDOT)*, which is defined as follows. *Days of Therapy (DOT)* is the number of different agents a patient receives on any given day of the admission. The *cumulative DOT (cDOT)* on day i is the cumulative sum of DOT from day 1 through day i. We also use the term *Days Since Admission (DSA)* to mean the number of days since the admission date (including the admission date). Note that when DSA equals LOS for a patient, the patient is either discharged or deceased.

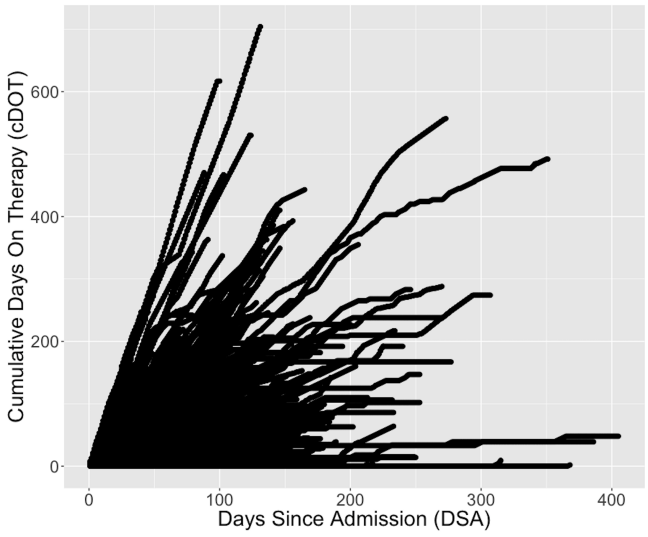


Figure 1: A scatterplot of patient trajectory data shown as a distribution of cumulative Days-of-Therapy (cDOT) values as a function of Day-Since-Admission (DSA). A single patient's trajectory of points is represented as a series of dots from DSA 1 to the last day of the patient's admission.

In Figure 1, we show a simple scatterplot of patients' trajectories that we obtained from the DASON database (see Section 4 for more details). It shows the distribution of cDOT values (performance, on y-axis) as a function of DSA (time, on x-axis). This scatterplot, while informative in its own to show the diversity of cDOT values, could become easily overwhelming for decoding or identifying any hidden patterns or substructures, particularly for large data sets containing millions of patients. Nevertheless we show this scatterplot to illustrate the simplistic view of data that it presents.

Hypothesis: We used the following two-part working hypothesis to guide our study in understanding antimicrobial exposure for inpatients:

- cDOT is responsive to a combination of temporal, spatial and treatment-based factors, although to varying degrees; and
- there can be significant variability across different (hidden) segments of the patient population in the way cDOT is correlated to these factors.

In other words, a patient's antimicrobial exposure is a combined function of time (i.e., their respective length of stay (LOS)), and is also potentially influenced by spatial attributes such as the units and hospitals they receive treatment in. Furthermore, we hypothesize that the type of antibiotic drug agents a patient receives in the earlier stages of their stay could influence the type of agents they receive in later stages of their stay.

Ideally, we would like to construct a robust mathematical model (or models) to describe how the antimicrobial exposure is a function of all the above factors. However, such a model construction is likely to require a significant and complex effort; instead in this paper, we focus on obtaining information from the data (of patient trajectories) that is already available, in order to guide future model construction efforts in a data-guided manner.

The second part of our hypothesis provides a way to contextualize the level of influence, as we expect variability in cDOT responses among different patient subgroups (or subpopulations). These subpopulations are not necessarily known a priori (i.e., they are hidden) and they need to be discovered as part of the analysis.

3.2 A Visual Analytics Framework

To test our working hypothesis, we implemented an unsupervised approach that has its principles rooted in the mathematical field of topological data analysis (TDA). Algebraic topology is the branch of mathematics dealing with the shape and connectivity of spaces [1, 14]. There are multiple important properties of topology that make it particularly effective for extracting structural features from large, high-dimensional data sets. First, topology studies shapes in a *coordinate-free way*, which enables comparison among data sets from diverse sources or coordinate systems. Second, topological constructions are *not sensitive to small changes in data*, and robust against noise. Third, topology works with *compressed representations* of spaces in the form of *simplicial complexes* (or triangulations) [14], which preserve information relevant to how points are connected. Compared to more traditional techniques such as principal component analysis, multidimensional scaling, and cluster analysis, topological methods are known to be more sensitive to both large and small scale patterns [11].

Our approach is unsupervised in that no prior information or models are assumed and that the approach makes its inferences entirely based on the data. However, we wish to point out that the inferences made by the TDA approach do not necessarily imply causality. They should be viewed as identifying *generalized* correlations between variables across the spectrum of a heterogeneous population—there is increased variability in the degrees of the correlations across the population. Such generalized correlations could not be identified by direct application of traditional data analysis techniques.

In this paper, we present an implementation for analyzing patient trajectory data sets using the `Hyppo-X` framework [8, 9], which is an implementation of the Mapper algorithm [17]. `Hyppo-X` is a computational tool for modeling and exploring multidimensional data where one set of (continuous) variables $f = \{f_1, f_2, \ldots f_k\}$ can be modeled as "filters" to study their impact on a target performance (continuous) variable g. In the context of our application, each f_i

variable can be any of our potentially influential variables (temporal, spatial, treatment-based), while the performance variable g is chosen as cDOT. We now elaborate the use of Hyppo-X framework in the context of our application.

Input: The input is a set of n points where each "point" is a unique combination of [patient id, hospital id, admission id, DSA]. These are the key fields that define a patient record (or a point) in our analysis. In addition, each point also has a set of attributes including, but not limited to, cDOT, NHSNunit-id, the agents that the patient was administered on that DSA, etc. Figure 2a shows an example input distribution of points (just for illustration purposes). Note that a trajectory can be defined by following the trail of points for a patient through the DSA interval [1, LOS]. As the LOS is different for each patient, the different trajectories are expected to be of varying lengths.

Output: Hyppo-X takes a set of input points as defined above and outputs a compact visual representation of the points, grouped into clusters that are connected via inter-cluster edges, that summarily shows the evolution of points along the particular dimension (or dimensions) chosen by the user.

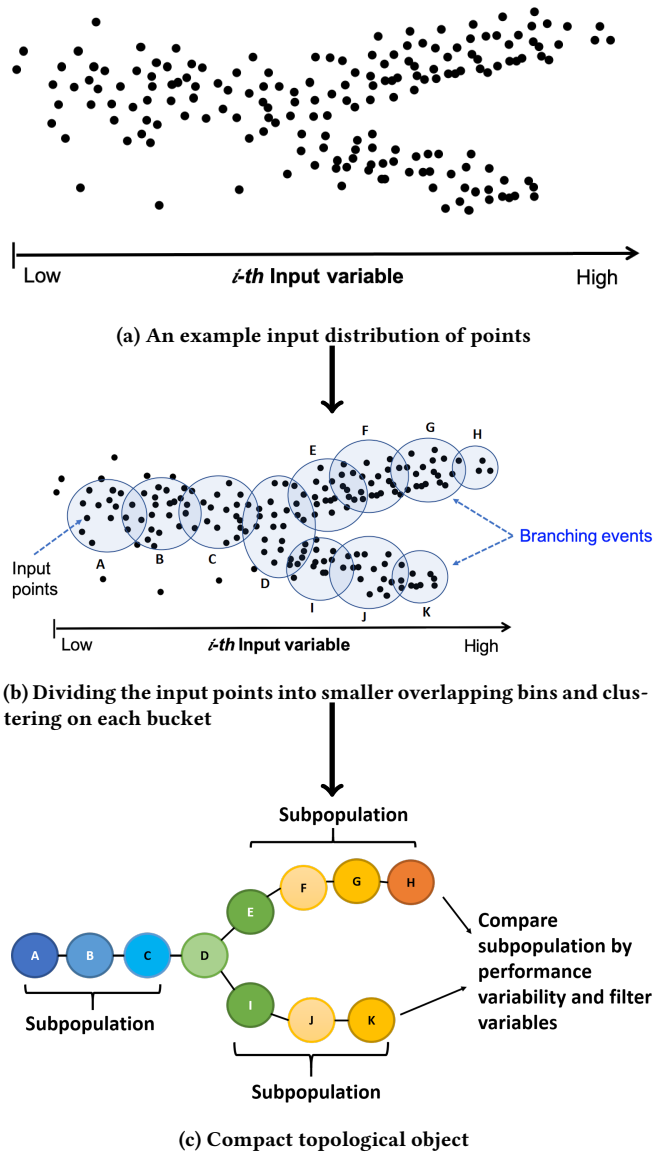
Algorithm: Let X denote the set of points, and f_z denote a particular dimension that we would like to use as a "filter" to view the set of points. Intuitively, a filter can be thought of as a variable of interest (e.g., DSA) that we would like to use as a "lens" through which we would like to view the entire distribution of points.

Given X and a filter f_z, the goal is to generate a graph-like representation of clusters, where each node in the graph is a cluster of points, and an edge exists between any two nodes if the corresponding two clusters intersect in points. Here, a cluster is a subset of points in X that show similar cDOT performance (i.e., have highly similar cDOT values) under a certain interval of the filter variable (e.g., DSA 5 through DSA 10).

Intuitively, each cluster represents a set of patient records that show similar cDOT values observed around the same DSA interval; and an edge exists between two nodes in our graph if the corresponding two clusters share at least one patient record in common. This representation allows us to track the progression of patient records as their trajectories evolve in time and cDOT performance.

Here is a more elaborate explanation of how we compute the clusters. We slide a sequence of fixed length overlapping windows over the filter variable, from the minimum DSA (=1) to the maximum DSA (=maximum over all the LOS). For experiments we used a window length of 5, and we allow adjacent windows to overlap by one day (i.e., last day of window i overlaps with the first day of window $i + 1$). Next, the set of points that fall into each window is organized into a bin (one bin for each window). We then run the density-based clustering algorithm DBSCAN [3] to cluster all points in a bin by their cDOT values. This clustering step produces a partitioning of points within each bin, i.e., no two clusters belonging to the same bin share any point in common. If a region is sufficiently dense, we make a cluster as shown in Figure 2b. Furthermore, because we allow the windows to overlap, we generate trails of overlapping clusters as shown.

Subsequently, a graph is generated, where every cluster is represented as a node, and an edge is drawn between any two nodes where the respective clusters share at least one point in common. Note that by construction, an edge can exist only between clusters



(a) An example input distribution of points

(b) Dividing the input points into smaller overlapping bins and clustering on each bucket

(c) Compact topological object

Figure 2: Generating topological object from a point cloud.

originating between two adjacent bins. We refer to the resulting graph as a *topological object* (simplicial complex, to be precise) as shown in Figure 2c. If three clusters share points, the object includes the triangle connecting the corresponding three nodes. In this work, we limit our attention to the vertices and edges in the topological object, i.e., its graph. This graph is a compact representation of the set of input points, and allows one to efficiently visualize a large collection of patients.

3.2.1 **Feature extraction:** We can extract features as a structural property of the topological object, which in turn help to generate hypotheses. One such structural feature is a "flare" that represents branching phenomenon in the topological object. We now describe

the structure of a flare; an algorithm for detection of flares was presented in our previous work [9].

A flare is a combination of a stem, branching node, and branches. A *stem* in a flare is a simple path that ends at a branching node. A *branching node* is a node that has at least two outgoing edges. Finally, a *branch* is a simple path that starts at a branching node and ends at either another branching node or a terminal node (zero out-degree node). For instance, in Figure 2c the nodes labeled $[A, B, C, D]$ refer to a stem. The node D is a branching node with two branches—one covering the path with nodes labeled $[D, E, F, G, H]$, and the other covering the path with nodes labeled $[D, I, J, K]$. The topological object contains a single flare here.

Branching phenomena as captured by a flare help us understand divergent behavior of two (or more) subpopulations covered by the branches. The comparative analysis of two divergent subpopulations could help us formulate and subsequently test plausible hypotheses pertaining to distinct behavior of hidden subpopulations of a larger population.

3.3 Software

We implemented the project in C++, PHP, and D3 (for visualization). The library generates graph objects in the JSON format for analysis. Our framework is publicly available and can be accessed as part of the `Hyppo-X` open source software kit [8].

4 EXPERIMENTAL RESULTS

4.1 Data

We used the Duke Antimicrobial Stewardship Outreach Network (DASON) database [2], which comprises of 25 community hospitals with full inpatient data. DASON contains detailed electronic medication administration records (eMAR) for antimicrobials, patient movement data (bed flow), demographics, and billing data. It includes information for millions of admissions, but we excluded the records for outpatients when preparing our final dataset. Also for calculating DOT, we counted only the antibacterial agents. We imposed some other constraints as well, e.g., removing null values, narrowing the dataset in between a specific range of dates, and so on. Table 1 provides a brief summary of the final data set that we used in all our analysis.

Table 1: Summary of the data set

Number of hospitals	25
Number of hospital unit-categories	9
Number of distinct patient-admission records	349, 610
Number of adult patients	334, 207
Number of male patients	148, 540
Number of female patients	201, 052
Average LOS per admission	7 days
Longest LOS	405 days
Number of antibacterials used	66
Most used antibacterial	Vancomycin
Average DOT per admission	6
Number of agent ranks	4
Most used agent rank	rank 3

4.2 Experimental Evaluation

We ran the Hyppo-X framework on our hospital data set using Days Since Admission (DSA) as a single filter function with bin size of 5 days. Figure 3 shows the static snapshot of the topological object output by our framework[1]. The clusters appear left to right in an increasing order of their mean DSA values. The label within each cluster node shows the mean cDOT value for the inpatients in the corresponding DSA interval. We can see that the branching phenomena starts to appear around day 70–80. This observation suggests that there is little divergence among the patients during the early part of their stay in the hospitals. However, the cluster size becomes smaller along the way, and branches with higher cDOT values start to emerge for longer term patients. This structure is expected because increasingly more patients are discharged with time.

We now analyze the distribution of patient clusters at different stages of their trajectories by showing each cluster as a pie-chart within it, based on different patient record attributes. The attributes we use to analyze (one at a time) include (but are not limited to): distribution of hospital units within clusters (Section 4.2.1), antibiotic agent ranks used within clusters (Section 4.2.2), and hospital-specific analysis (Section 4.2.3).

4.2.1 Analysis based on unit category for patient clusters. There are 42 hospital units in our data set. These units can be grouped into 9 categories. In Figure 4, the pie-charts show the distribution of the hospital unit categories within each cluster.

We make the following observations based on Figure 4:

- The majority of the clusters are dominated by patients in the adult medical/surgical ward (shown in red), followed by adult critical care (shown in green). However, these patient clusters correspond mostly to the low cDOT branches (see corresponding clusters in Figure 3), suggesting a relatively low use of antibiotics for these patient groups.
- The composition of clusters (by unit categories) start to change in the later branches of the object (with DSA values of 100 or more). In fact, on one dominant set of branches around DSA 100, we see a more even distribution among adult critical care, pediatric critical care, and Hematology/Oncology/Transplant wards. These clusters also see a relative spike in their cDOT values (see corresponding clusters in Figure 3).
- Another interesting observation is that there is a distinctive set of cluster branches in around DSA 130 and above, that also see an increase in their cDOT values. This set of cluster branches is dominated with patients from the neonatal unit (shown in light green). In addition, we see a divergence in cDOT usage even among this small group of neonatal unit patients—with some branches receiving a higher cDOT values than the others.

Collectively, these observations suggest that antibiotic use does *not* necessarily show a linear increase with time. Instead, different patient groups receiving treatment in different units show spikes in their antibiotic use at different intervals of their hospital stay. Furthermore, not all units see a comparable use of antibiotics—for instance, adult medical/surgical ward is frequently occurring but

[1] Note that in the actual tool, all topological objects allow interactive visualization.

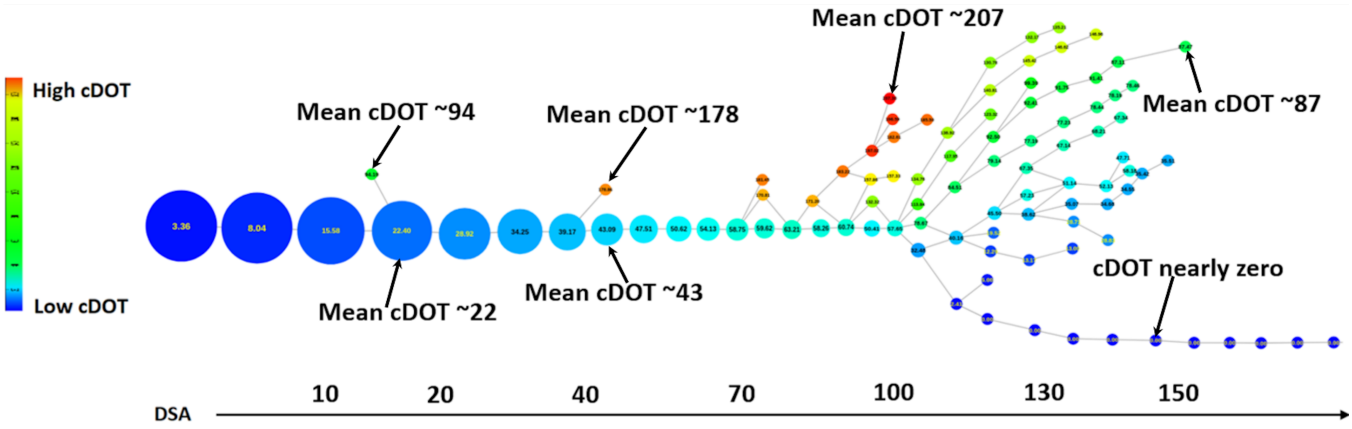


Figure 3: The topological object constructed using Days Since Admission (DSA) as a single filter function and cumulative Days On Therapy (cDOT) as the clustering attribute. The horizontal color bar indicates the gradient of DSA value from left (low) to right (hight). The node coloring and the labels are both based on the mean cDOT value for that cluster. The node color spectrum is from blue (low) to red (high). Each cluster is defined using an interval of 5 days. Hence the first cluster from left represents the first five days of the patients admitted in the hospital, the second cluster represents 5^{th}–10^{th} days of the patients' stay, and so on.

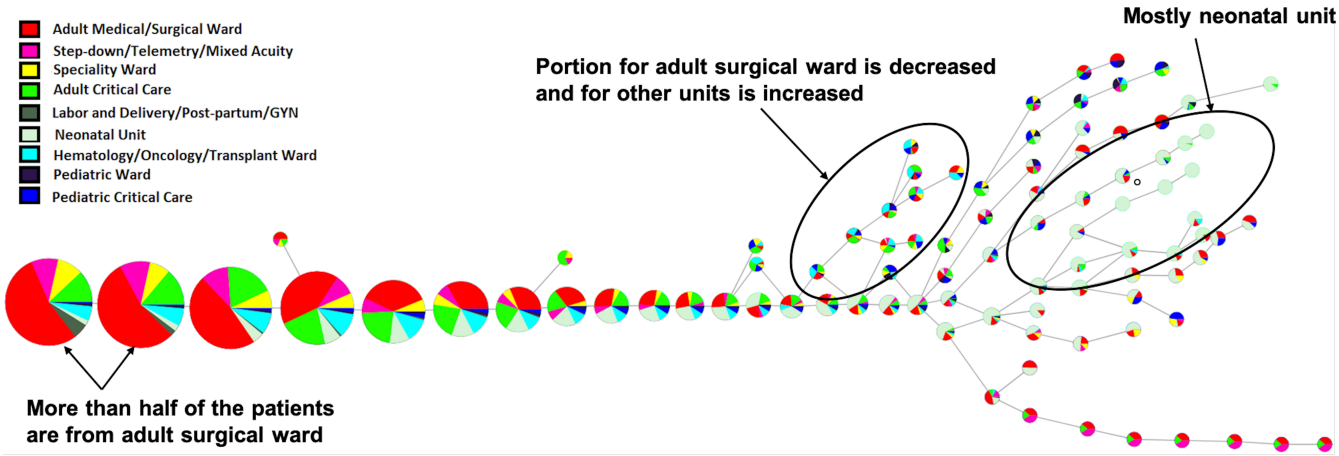


Figure 4: The distribution of hospital unit categories shown as a pie-chart within each cluster.

receives lower antibiotics; whereas pediatric ward or some segment of neonatal unit populations are rarer but receive higher antibiotics. There are also units that are both rare and are exposed to lower antibiotics (e.g., labor and delivery/post-partum/GYN). Finally, there is also a cDOT divergence *within* the same unit category—in particular, patient groups in the neonatal ward.

4.2.2 Analysis based on agent rank on patient clusters. There are a total of 66 antibacterial agents used on patients in the DASON data. We can rank and group these agents into four groups—from rank 1 through rank 4—roughly in order of their type/target microbial coverage. This ranking also reflects a rough ordering based on the agent severity (with 1 being low to 4 being high).

Using this ranking scheme, we computed the distribution of agent ranks used within each cluster. This distribution is as per the agents used by the patients in a given cluster (within the DSA range represented by that cluster). In addition, there were many days when a patient did *not* receive any agent. To capture such cases, we introduced a separate "No agent usage" rank category. Figure 5 shows the distribution of agent ranks within each cluster. We make the following observations based on this figure:

- The most dominant category is the "No agent usage" category across the range of clusters. However, in the initial days of stay (DSA range 1 through 60–70) this is not necessarily true (i.e., other agent ranks are visible).
- Among the agents used, rank 3 agents appear most frequent (shown in blue), followed by rank 1 (shown in yellow), and subsequently by rank 2 (shown in cyan).

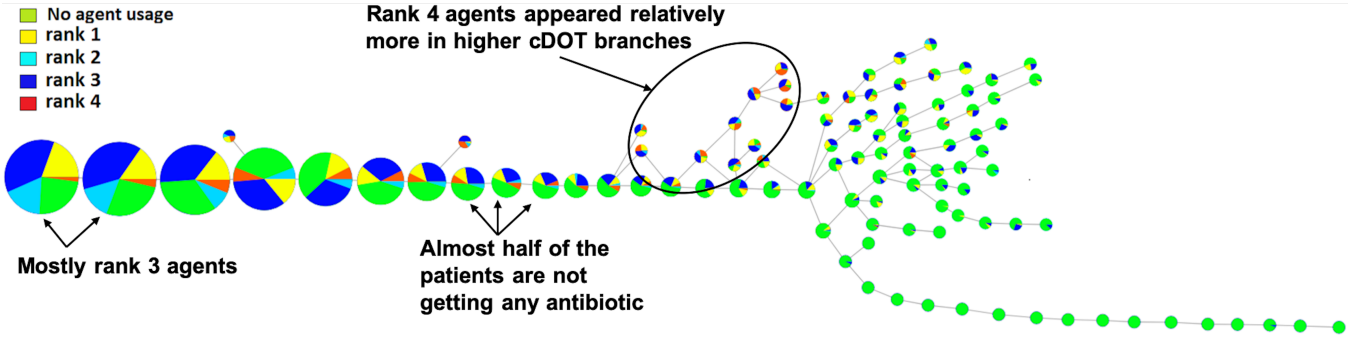


Figure 5: The pie chart in each node is representing the percentage usage of each agent group on that particular interval of time.

- Rank 4 agents appear rarely (represented by red) but they also generally appear in the branches with the higher cDOT values. This observation probably suggests that use of this agent is reserved typically for patients with worsening health conditions.

Note here that the use or non-use of an antibiotic agent (rank) could potentially be a matter of preference or practice protocol across different hospitals. In the following section, we analyze their impact across different hospitals.

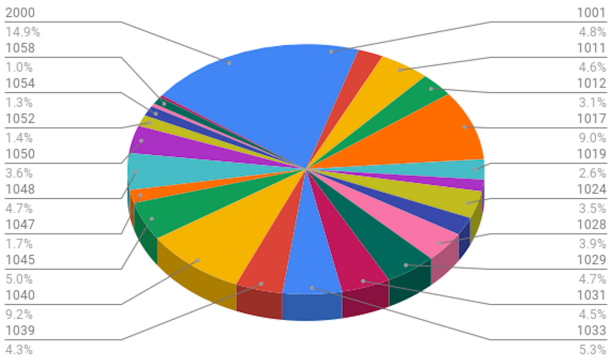


Figure 6: Number of records per hospital. The Duke hospital (id: 2000) has the largest number of records.

4.2.3 Rank based analysis on specific hospital. Patient data from a total of 25 hospitals are represented in the DASON data set. However, the Duke medical hospital (hospital id: 2000) is the dominant contributor accounting for almost 15% of the unique patient records—as shown in Figure 6. To elucidate any potential differences in antibiotic use across these different hospitals, we performed two studies—one by considering records only from the Duke hospital (id: 2000), and another by considering records only from the remaining 24 hospitals. The resulting objects are shown in Figure 7. Even though the pie-charts in the clusters are shown by their agent rank distributions, we also compare information that is contained in the general structure of these two objects.

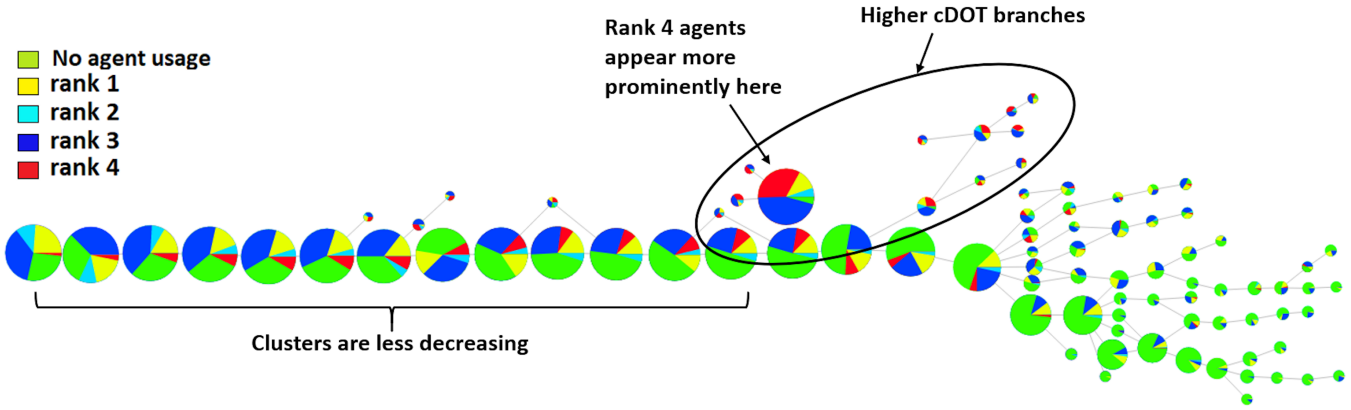
We make the following observations based on Figure 7:

- The sizes of the clusters stay roughly uniform over the first 100 days along the main stem of the topological object for Duke hospital, whereas the sizes rapidly shrink for the other hospitals in the same period. This suggests that the patients tend to stay longer at the Duke hospital, whereas in the community hospitals, patients either leave because they recover or they leave to get transferred to a larger hospital (such as Duke).
- Patients in the Duke Hospital are more likely (than ones in the other hospitals) to receive some antibiotic at least once during their stay.
- The use of agent rank 4 is relatively more frequent at the Duke hospital than for the other hospitals.
- Even though these two objects were constructed individually, the general topological structure (i.e., overall shape) is roughly comparable, suggesting we have similar branching/divergence attributes between the two classes of hospitals (Duke vs. non-Duke).

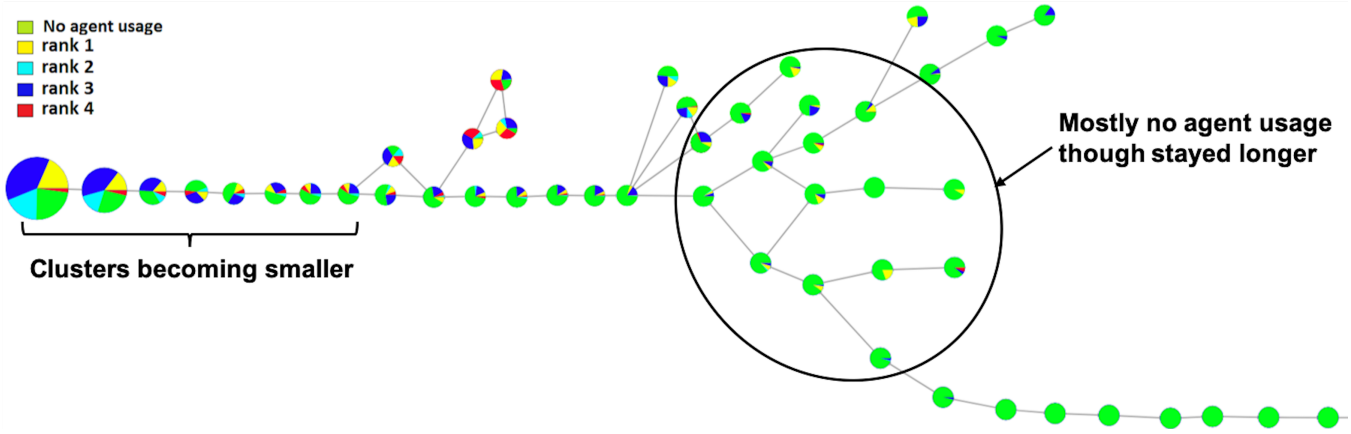
4.3 Interesting features/flares

So far, we have described observations on the topological object without necessarily examining its branching structure in detail. In order to more thoroughly examine the branching structure within different parts of the object (i.e., different subpopulations), we applied our flare detection algorithm (Section 3.2.1) to the DASON data. Recall that a flare is a structural features comprising of a stem region that ends at a branching node, and is subsequently followed by a number of child branches. For the purpose of our study, we used the hospital id attribute of the dataset to identify the coverage of a flare. The coverage of a flare specifies the boundary to which we can extend a given flare. This is done in order to ensure that we recover a meaningful branch which covers data points from the same subpopulation.

Figure 8 shows the two most interesting flares detected by our approach (shown in blue and red arcs). Note that our tool computes a score for each flare and outputs them in decreasing order. We make the following observations based on the detected flares:



(a) Rank based analysis on the data set only from Duke Hospital.



(b) Rank based analysis on the data set excluding Duke Hospital.

Figure 7: A comparison of rank based analysis on Duke (top) vs Non-Duke (bottom) Hospitals.

(1) From DSA 1 to DSA 80, there is little divergence in the cDOT values of the patient clusters (with a few exceptions), and this is shown by the long stem of the blue flare.
(2) This behavior changes around DSA 80. A group of patients were treated with higher dose of antibiotics compared to the remaining group (Figure 8(A)). The branching node (shown with thick border) in the blue flare represents this branching event. This branching event essentially serves to bifurcate patients in the Hematology/Oncology/Transplant or the Pediatric wards into two subgroups as shown in Figure 8(C)—those for whom cDOT increased (higher branches) and those for whom it did not, along the lower branches.
(3) The other flare (shown in red arcs), with a branching occurring around DSA 100, shows a further split in the population between the neonatal branches (lower) vs. non-neonatal branches (higher).
(4) In terms of agent rank usage, we see that it is the subpopulation corresponding to the first flare (blue) that is exposed to agent rank 4 (see Figure 8(B)). This subpopulation corresponds to patients mostly in either the Adult Medical/Surgical Ward or the Hematology/Oncology/Transplant Ward (see Figure 8(C)).

In summary, our approach was able to identify in an unsupervised manner the major branching events in the data. Further, the analysis presented above shows which subgroups within the larger patient population are more prevalent in those branches.

5 CONCLUSIONS

The use of topological data analysis for biomedical informatics applications is relatively new. This technique has a potential to represent complex data in compact and visually-friendly formats. In this paper, we have presented a visual analytics framework powered by topological data analysis for analyzing patient trajectories, and applied it to the concrete application of exploring patient trajectories following the use of antibiotics in hospitals. First, the framework allowed a clustering of patient trajectories (by their antibiotic use, or cDOT) so that they can be concisely viewed along the temporal dimension. Subsequently, we used multiple attributes such as hospital units or agent ranks to analyze and observe patterns (both conserved and discrepant substructures) in these clustered trajectories.

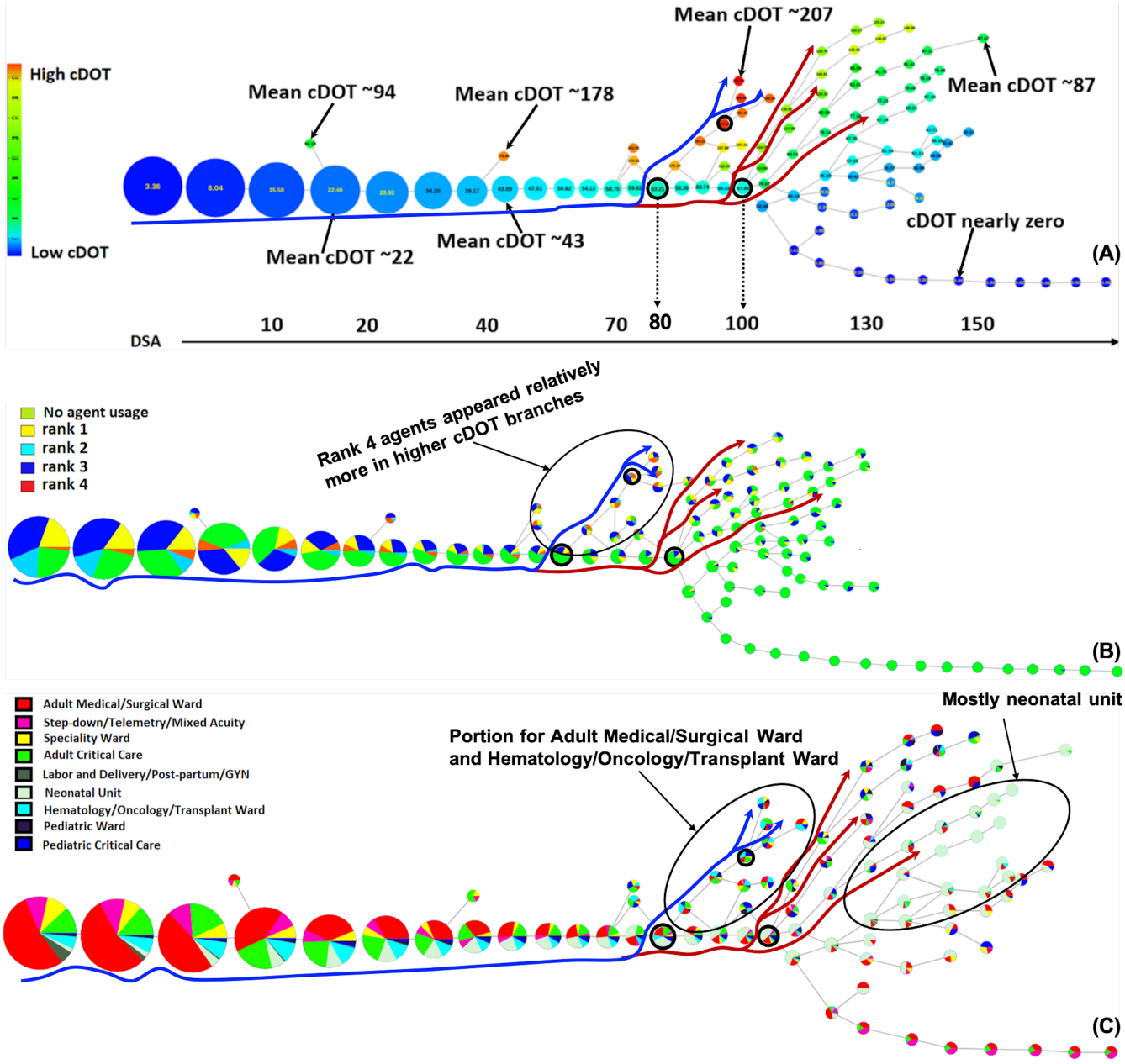


Figure 8: Topological object constructed using DSA as a single filter function (shown earlier in Figure 3), now also showing the interesting flares detected by our method. The nodes are arranged from left to right with chronological order of mean DSA values. (A) Each cluster colored by its mean cDOT, with branches showing different degree of uses. (B) Each cluster (node) of the topological object is rendered as a pie-chart showing the distribution of their five antibiotic classes. (C) Each cluster (node) of the topological object is rendered as a pie-chart showing the distribution of their nine hospital unit classes. Long arcs of different colors show interesting flares, and the corresponding branching nodes are identified with bold border. The blue flare was ranked as the most interesting flare.

In contrast to traditional approaches using dense scatterplots (as shown in Figure 1), our approach provides the new capability to analyze and examine patient trajectories in a detailed and visually interactive manner. Further, our analysis by different attributes revealed a more nuanced picture of how antibiotics are used across the patient population and across hospitals (and their units). In summary, we not only find differing propensities for use

of antibiotics within certain hospital units, but we also find similar propensities across hospitals. Furthermore, we observe divergence within patient groups (e.g., neonatal) on how antibiotics are used. These observations are directly inferred from the data in an unsupervised manner, and could in turn inform future construction of more robust models in this space.

Future research directions include (but are not limited to) the following. In addition to the attributes used, we plan to explore using other variables in our framework to study antibiotic use in hospitals including patients' Elixhauser score (which gives an indication of comorbidities in a patient), disease diagnostic codes (associated with each admission), and others. These variables could collectively throw more light into the context under which a patient receives treatment in a hospital.

The observations made in this work also open new questions about what makes a patient more susceptible toward antibiotic exposure in hospitals, and about whether there is a way to build predictive/probabilistic models based on training data obtained from these trajectories. Also, more work is needed to understand and better characterize the structural properties of the topological objects created for different hospitals. In particular, comparing and contrasting them can help us better understand similar and discrepant practices across those healthcare locations and also help us devise consistent and standardized procedures toward improving antibiotic stewardship.

ACKNOWLEDGMENTS

This research was supported in parts by a U.S. Center for Disease Control and Prevention (CDC) contract to Washington State University and Duke University, and by U.S. National Science Foundation (NSF) grant DBI 1661348.

REFERENCES

[1] Carlsson, G. Topology and data. *Bulletin of the American Mathematical Society 46*, 2 (Jan. 2009), 255–308.

[2] Duke University School of Medicine. Duke Antimicrobial Stewardship Outreach Network (DASON). https://dason.medicine.duke.edu/, 2019.

[3] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD* (1996), vol. 96, pp. 226–231.

[4] Giannoula, A., Gutierrez-Sacristán, A., Bravo, Á., Sanz, F., and Furlong, L. I. Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study. *Scientific reports 8*, 1 (2018), 4216.

[5] Goldstein, B. A., Navar, A. M., Pencina, M. J., and Ioannidis, J. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association 24*, 1 (2017), 198–208.

[6] Häyrinen, K., Saranto, K., and Nykänen, P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics 77*, 5 (2008), 291–304.

[7] Jensen, A. B., Moseley, P. L., Oprea, T. I., Ellesøe, S. G., Eriksson, R., Schmock, H., Jensen, P. B., Jensen, L. J., and Brunak, S. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature communications 5* (2014), 4022.

[8] Kamruzzaman, M. HYPPO-X: A software library for visual analytics on complex high dimensional data. https://xperthut.github.io/HYPPO-X, 2019.

[9] Kamruzzaman, M., Kalyanaraman, A., and Krishnamoorthy, B. Detecting divergent subpopulations in phenomics data using interesting flares. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (2018), ACM, pp. 155–164.

[10] Li, L., Cheng, W.-Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E. P., and Dudley, J. T. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine 7*, 311 (2015), 311ra174–311ra174.

[11] Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J. G., and Carlsson, G. Extracting insights from the shape of complex data using topology. *Scientific Reports 3*, 1236 (2013).

[12] Mannhardt, F., and Blinde, D. Analyzing the trajectories of patients with sepsis using process mining. In *RADAR+ EMISA@ CAiSE* (2017), pp. 72–80.

[13] Mitchell, B. G., Ferguson, J. K., Anderson, M., Sear, J., and Barnett, A. Length of stay and mortality associated with healthcare-associated urinary tract infections: a multi-state model. *Journal of Hospital Infection 93*, 1 (2016), 92–99.

[14] Munkres, J. R. *Elements of Algebraic Topology*. Addison–Wesley Publishing Company, Menlo Park, 1984.

[15] Nekkab, N., Astagneau, P., Temime, L., and Crepey, P. Spread of hospital-acquired infections: A comparison of healthcare networks. *PLoS computational biology 13*, 8 (2017), e1005666.

[16] Nicolau, M., Levine, A. J., and Carlsson, G. 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), 7265–7270.

[17] Singh, G., Mémoli, F., and Carlsson, G. E. Topological methods for the analysis of high dimensional data sets and 3d object recognition. pp. 91–100. 00045.