KDD2019

25TH ACM
SIGKDD
CONFERENCE
ON KNOWLEDGE DISCOVERY
AND DATA MINING

ANCHORAGE, ALASKA
AUGUST 4–8, 2019
Dena'ina Convention Center and
William Egan Convention Center

epiDAMIK: Epidemiology meets Data Mining and Knowledge discovery

Workshop held in conjuction with ACM SIGKDD 2019

Anchorage, Alaska - USA. August 5, 2019

epiDAMIK
@KDD2019

Workshop Proceedings

Editors: B. Aditya Prakash, Anil Vullikanti, Shweta Bansal, Adam Sadelik

Proceedings of the ACM SIGKDD Workshop on Epidemiology meets Data Mining and Knowledge Discovery (epiDAMIK)

These proceedings are not included in the ACM Digital Library.

epiDAMIK'19, August 5, 2019, Anchorage, Alaska, USA.

ACM SIGKDD Workshop on Epidemiology meets Data Mining and Knowledge Discovery (epiDAMIK)

Organizers:

B. Aditya Prakash (Virginia Tech)
Anil Vullikanti (University of Virginia)
Shweta Bansal (George Washington University)
Adam Sadelik (Google)

Webmaster:

Bijaya Adhikari (Virginia Tech)

Preface

With increasing globalization, urbanization, and ecological pressures, the threat of devastating global pandemics becomes more pronounced. The impact of Zika, MERS, and Ebola outbreaks over the past decade has strongly illustrated our enormous vulnerability to emerging infectious diseases. There is an urgent need to develop sound theoretical principles and transformative computational approaches that will allow us to address the escalating threat of a future pandemic. Data mining and Knowledge discovery have an important role to play in this regard. Different aspects of infectious disease modeling, analysis and control have traditionally been studied within the confines of individual disciplines, such as mathematical epidemiology and public health, and data mining and machine learning. Coupled with increasing data generation across multiple domains (like electronic medical records and social media), there is a clear need for analyzing them to inform public health policies and outcomes. Recent advances in disease surveillance and forecasting, and initiatives such as the CDC Flu Challenge, have brought these disciplines closer––public health practitioners seek to use novel datasets and techniques whereas researchers from data mining and machine learning develop novel tools for solving many fundamental problems in the public health policy planning process. We believe the next stage of advances will result from closer collaborations between these two communities, which is the main objective of epiDAMIK. The workshop is also an integral part of the 'Health Day @ KDD', which brings together domain and machine learning experts to discuss challenges and trends in the healthcare industry as well as techniques and methodologies the machine learning community is using, and in process of developing to address these challenges.

The main program of epiDAMIK'19 consists of four papers that cover various aspects of data mining and public health. In addition there were two keynotes. All the papers were presented orally, and also during an interactive session joint with the KDD poster session as part of the Health Day. We sincerely thank the authors of the submissions and the attendees of the workshop. We also wish to thank the members of our program committee for their help in selecting a set of high-quality papers. Furthermore, we are very grateful to Ben Althouse and Elaine Nsoesie for engaging keynote presentations.

B. Aditya Prakash
Anil Vullikanti
Shweta Bansal
Adam Sadelik

Arlington, August 2019

Table of Contents

Invited Talks

Research Papers

Invited Talk

Novel data streams (NDS) for surveillance

Benjamin Althouse
Principal Scientist, Co-Chair of the Epidemiology team
Institute for Disease Modeling
balthouse@idmod.org

Abstract:

Novel data streams (NDS), such as web search data or social media updates, hold promise for enhancing the capabilities of public health surveillance. In this talk I will explore several case-studies of how NDS can be used for infectious disease surveillance; understanding how individuals seek information about health behaviors such as depression and smoking cessation; "organic advocacy" wherein celebrities or other viral announcements affect individual search behavior; "crowd diagnoses" of sexually transmitted diseases on Reddit; and novel surveillance of bed bug infestations. Through these disparate examples of the uses of NDS, I highlight the necessities of accurate model building and thorough and appropriate interpretation of surveillance using NDS.

Bio:

Ben Althouse is a Principal Scientist and Co-chair of the Epidemiology team at the Institute for Disease Modeling where he explores pneumococcal pneumonia vaccines, the transmission dynamics of respiratory pathogens, and the role of complex human contact structures on disease transmission. He was an Omidyar Fellow at the Santa Fe Institute, holds a PhD in Epidemiology and a Master of Science in Biostatistics from the Johns Hopkins Bloomberg School of Public Health where he was awarded an NSF Graduate Research Fellowship, and holds Bachelor of Science degrees in Mathematics and Biochemistry from the University of Washington. His previous work has included mathematical modeling of sylvatic dengue virus transmission in nonhuman primates in Senegal, examining the role of antimicrobial use on the evolution of drug resistance, using Twitter as a model system of co-infection dynamics, and using novel data sources (such as Google searches, Twitter, and Wikipedia article views) for population-level surveillance of infectious and chronic diseases. Ben is an Affiliate Faculty member in the Department of Biology at New Mexico State University, Las Cruces, and an Affiliate Assistant Professor at the Information School at the University of Washington.

Invited Talk

Non-traditional Approaches to Public Health Surveillance

Elaine Nsoesie
Assistant Professor
School of Public Health
Boston University.
onelaine@bu.edu

Abstract:

Data from a variety of sources, including social media, e-commerce websites and remote sensing, offer unique opportunities for studying and addressing problems in public health. In this talk, we will present examples on the use of data from a variety of sources for public health surveillance. We will also discuss the biases inherent in these datasets and potential implications on the public's health.

Bio:

Dr. Nsoesie is an Assistant Professor of Global Health at Boston University (BU). She is also a BU Data Science Faculty Fellow as part of the BU Data Science Initiative at the Hariri Institute for Computing, and a Data and Innovation Fellow at The Directorate of Science, Technology and Innovation (DSTI) in the Office of the President in Sierra Leone. Dr. Nsoesie applies data science methodologies to global health problems, using digital data and technology to improve health, particularly in the realm of surveillance of chronic and infectious diseases. She completed her PhD in Computational Epidemiology from the Genetics, Bioinformatics and Computational Biology program at Virginia Tech. She also have an MS in Statistics and a BS in Mathematics. She has written for NPR, The Conversation, Public Health Post and Quartz. Dr. Nsoesie was born and raised in Cameroon.

Improved Automatic Pharmacovigilance: An Enhancement to the MedWatcher Social System for Monitoring Adverse Events

Andre T. Nguyen
Booz Allen Hamilton
Nguyen_Andre@bah.com

Julia Lien
Booz Allen Hamilton
Lien_Julia@bah.com

Edward Raff
Booz Allen Hamilton
Raff_Edward@bah.com

Sumiko R. Mekaru
Booz Allen Hamilton
Mekaru_Sumiko@bah.com

ABSTRACT

Traditional pharmacovigilance systems rely on adverse event reports received by regulatory authorities such as the United States Food and Drug Administration (FDA). These traditional systems suffer from underreporting and are not timely due to their reliance on third-party sentinels. To address these issues, the MedWatcher Social system for monitoring adverse events through automated processing of digital social media data and crowdsourcing was launched in 2012 by Boston Children's Hospital and the FDA. The system is rooted in the well-established FDA MedWatch system.

MedWatcher Social uses an indicator score approach to identify adverse events. This study evaluates the MedWatcher Social adverse event classifier's performance on Twitter data and proposes an enhancement to the indicator score method that results in improved adverse event identification.

Our research suggests that automatic pharmacovigilance systems using the original indicator score approach should be updated. Careful consideration of modeling assumptions is critical when designing algorithms for computational epidemiology, and algorithms should be regularly reevaluated to identify enhancements and to remedy concept drift.

ACM Reference Format:
Andre T. Nguyen, Julia Lien, Edward Raff, and Sumiko R. Mekaru. 2019. Improved Automatic Pharmacovigilance: An Enhancement to the MedWatcher Social System for Monitoring Adverse Events. In *Proceedings of epiDAMIK: Epidemiology meets Data Mining and Knowledge Discovery Workshop at KDD (epiDAMIK @ KDD'19).* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

Pharmacovigilance, also known as drug safety monitoring, is the pharmacological science relating to the detection, assessment, understanding, and prevention of adverse effects and other possible drug-related problems [19]. Although the safety of a drug is assessed through clinical trials before approval by regulatory authorities, the short timescale and the limited number of participants involved prevent clinical trials from comprehensively uncovering all possible adverse effects and drug interactions. Indeed, extremely rare adverse events are expected to be missed barring impractically massive clinical trials. As a result, continuous drug surveillance via pharmacovigilance is critical for consumer safety.

Traditional pharmacovigilance systems rely on spontaneous adverse event (AE) reports received by regulatory authorities such as the United States Food and Drug Administration (FDA). These traditional systems suffer from underreporting and lack of timeliness due to their reliance on third-party sentinels which are affected by lack of adverse event mentions from patients experiencing these events and then lack of follow through by medical staff who are often overtasked. The availability of adverse event data received by traditional government pharmacovigilance systems has a known lag time of 6 to 12 months [3]. A 2010 study conducted by the United States Department of Health and Human Services Office of the Inspector General found that 27 percent of hospitalized Medicare beneficiaries in October 2008 experienced adverse events during their hospital stays that required treatment, half of which resulted in prolonged hospitalization, required life-sustaining intervention, caused permanent disability, or resulted in death [7]. A follow up study found that 86 percent of the adverse events were not reported to a traditional government pharmacovigilance system [8]. To address the issues with traditional government pharmacovigilance systems, multiple automatic pharmacovigilance techniques using social media data have been suggested [18]. Drug surveillance systems based on social listening and intelligent automation are valuable and complement traditional systems because they monitor a mostly different population from that monitored by traditional systems and because they do not suffer from a data availability lag.

The MedWatcher Social system for monitoring adverse events through automated processing of digital social media data and crowdsourcing was launched in 2012 by Boston Children's Hospital and the FDA [5, 6, 12, 13]. The system is rooted in the well-established FDA MedWatch system. Recent research in the field of digital disease detection has shown that computational epidemiology systems should be reevaluated frequently to ensure that the best methodologies are being used [16, 17]. In this paper, we evaluate the MedWatcher Social adverse event classifier's performance on Twitter data and propose an enhancement to the indicator score method that results in improved adverse event identification. We also show how a careful consideration of modeling assumptions is critical when designing algorithms for computational epidemiology. Additionally, the pharmacovigilance community has developed some of its own tools which have existing counterparts in the machine learning space. In many fields, the tools of machine learning often outperform community built tools. We hope to illustrate an example of how and why this is the case.

epiDAMIK @ KDD'19, August 5th, 2019, Anchorage, Alaska, USA
2019.

2 METHODS

2.1 Data

This study used English-language public Twitter posts collected by MedWatcher Social. The Twitter data was ingested from third-party data vendors, filtering for selected medical product names and synonyms. After ingestion, tweets were processed by a taxonomy-based tagger to extract symptoms and products associated with each tweet. The tagger used an expert-curated ontology that maps colloquial synonyms to formal names of products and symptoms. The symptoms for which a tagged product is used to treat, also known as indications, were removed from the extracted list of symptoms. Duplicated tweets, most commonly retweets, were then removed using a string comparison algorithm [5, 6, 12, 13].

2.2 Curation

The MedWatcher Social system defines an adverse event report as a post consisting of an identified medical product, a specific patient, and a physiological or cognitive symptom that is believed to be due to the product [6]. The primary objective is to identify posts that resemble an adverse event (proto-AE) containing product-symptom associations. The secondary objective is to label non-proto-AE posts as mentions or junk, where a mention is a post in which a drug was discussed and a junk post is a tweet that is not of interest for pharmacovigilance. For training and testing purposes, each post was labeled by a MedDRA certified curator as Proto-AE, Mention, or Junk. An example of each type post can be found in Table 1.

Table 1: Example posts.

Post Type	Example Post
Junk	I'm watching Baret Jackson auction: Commercials are for Viagra, diarrhoea, online dating & Arthritis. On that note I'm gonna attempt to...
Mention	the flu shot lady didn't give me a band-aid, i could bleed to death.
Proto-AE	arms are swollen, think im allergic to ibuprofen :(

We sampled 100 000 curated data points from the data ingested by MedWatcher Social from 2011-08-18 to 2017-11-20. This sample was then split into a training set of size 70 000 and a test set of size 30 000. The class distribution of the data is 64% Junk, 21% Proto-AE, and 15% Mention. Usually, mainly posts tagged as Proto-AE by the MedWatcher Social algorithm were curated, with the objective of removing false positives. The curated data pool is thus likely to be somewhat different from the overall ingested data pool as a good number of the posts labeled as Junk or Mention by a curator represent harder to classify examples. As a result, reported performance metrics in this study are likely to be lower than what would be observed in an actual deployment of the algorithms on all of the ingested but not fully curated data.

2.3 Feature Extraction

In order to represent the text data in a format amenable to machine learning algorithms, we transform the posts into numerical form as follows:

(1) Tokenization: Given a sequence of symbols, tokenization chops the sequence up into pieces called tokens. The set of all observed tokens across posts is called the vocabulary. We used word unigrams, bigrams, and trigrams as tokens for our study.
(2) Bag-of-Words: Bag-of-words is a representation model often used in natural language processing for simplifying text data. Bag-of-words ignores token order but keeps token duplicates. Not all information from word ordering is lost however as our use of bigrams and trigrams preserves local ordering information.
(3) Count Vectorization: Given a bag-of-words representation of a text document, a numerical representation of the document can be computed by counting the number of occurrences of vocabulary tokens in the document.

2.4 Spectral Embedding via Laplacian Eigenmaps

High-dimensional data such as text documents in count vector form are hard to visualize. Dimensionality reduction can be used to help visualize the structure of a dataset. Manifold learning is a nonlinear dimensionality reduction technique that assumes that high-dimensional data can be mapped to a lower-dimensional embedding that locally preserves distance relationships.

Spectral embedding using Laplacian eigenmaps is a technique for constructing the lower-dimensional nonlinear embedding using a graph that discretely approximates the lower-dimensional manifold [1, 2]. There are a few variations of the algorithm. The simplest to tune instantiation of the algorithm is as follows. Given k data points in $\mathbb{R}^d$, a weighted graph with k nodes is built with each node corresponding to a data point. A n nearest neighbors approach is used to connect the graph where nodes i and j are connected by an edge if data points i or j are among each other's n nearest neighbors. Let W be the adjacency matrix of the resulting graph, let D be a diagonal matrix whose entries are column or row sums of W, and let L be the Laplacian matrix $L = D - W$. The eigenvalues and eigenvectors for the generalized eigenvector problem are then computed: $L\boldsymbol{y} = \lambda D\boldsymbol{y}$. Let $\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{k-1}$ be the eigenvector solutions ordered from smallest to largest by eigenvalue. The projection of data point i under the embedding into a lower dimensional space $\mathbb{R}^m$ is $\left(\boldsymbol{y}_1(i), \ldots, \boldsymbol{y}_m(i)\right)$.

2.5 Multinomial Naïve Bayes

The multinomial naïve Bayes classifier is a generative classifier for multinomial distributed feature vectors such as those resulting from count vectorization. Generative modeling has the goal of explaining how to generate data by learning the class conditional distribution of the feature vectors $p(\boldsymbol{x}|y = c)$ and the class prior $p(y = c)$. The class posterior, the probability of a class given the data, can be computed from the class conditional and the class prior using Bayes rule:

$$p(y = c|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y = c)p(y = c)}{p(\boldsymbol{x})} \propto p(\boldsymbol{x}|y = c)p(y = c)$$

The evidence $p(\boldsymbol{x})$ does not depend on the class so can be ignored in practice. Naïve Bayes takes the simplest approach for specifying the class conditional distribution by assuming that features are conditionally independent given the class. In the case of the multinomial naïve Bayes classifier, the class conditional is a multinomial distribution with independent event probabilities. For D dimensional data, this is mathematically:

$$p(\boldsymbol{x}|y = c) = \prod_{d=1}^{D} p\left(x_d|y = c\right)$$

Inference for the multinomial naïve Bayes model consists of learning the class prior and the event probabilities that form the class conditional. Naïve Bayes has sometimes been inappropriately called a Bayesian probabilistic algorithm in the digital pharmacovigilance literature. It is important to note that the use of Bayes rule is not enough to make an algorithm statistically Bayesian. If, as is most commonly done in terms of naïve Bayes inference, the full posterior is not computed and a maximum likelihood estimate or maximum a posteriori estimate is used to fit naïve Bayes, then the algorithm is not Bayesian [9].

2.6 Logistic Regression

Logistic regression is a discriminative classifier. While generative classifiers such as naïve Bayes model the class posterior $p(y = c|\boldsymbol{x})$ indirectly by modeling the class conditional and the class prior, discriminative classifiers fit the class posterior directly. In the two-class case, logistic regression corresponds to the following model where $\boldsymbol{w}$ is a vector of parameters:

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \text{Bernoulli}(y|p(y = 1|\boldsymbol{x})) = \text{ Bernoulli }\left(y|\text{ sigmoid }\left(\boldsymbol{w}^T\boldsymbol{x}\right)\right)$$

Inference for logistic regression model consists of learning the model parameters $\boldsymbol{w}$. Logistic regression can be regularized by placing a prior on $\boldsymbol{w}$ or by adding a regularization term to the objective function. The two-class logistic regression model can be extended to a multi-class setting using a one versus rest scheme where a binary classifier is trained for each class separately. A second way to extend logistic regression to a multi-class setting is through the use of the softmax function, a multidimensional generalization of the logistic function.

2.7 Classification Using Indicator Scores

The MedWatcher Social system currently uses an indicator score approach for adverse event detection [13]. The indicator score technique uses a two-class naïve Bayes classifier that discriminates Proto-AE posts from Junk posts. The naïve Bayes probability estimate of belonging the Proto-AE class is used as the indicator score. If the indicator score is greater or equal to an upper threshold, then the post is classified as Proto-AE. If the indicator score is less or equal to a lower threshold, then the post is classified as Junk. If the indicator score is strictly between the lower and upper thresholds, then the post is classified as Mention.

Some studies have used a variant of the naïve Bayes classifier, called the Robinson classifier, that combines event probabilities using Fisher's method for combining p-values instead of conditional probability [12]. The justification for the use of Fisher's method has been that it does not assume independence [15]. This justification is somewhat unsatisfying as event probabilities are not p-values and, more importantly, the form of Fisher's method used in these studies is meant for independent tests [4]. The Robinson variant of the indicator score approach also includes a penalty adjustment for lack of symptom mentions in a post. For completeness, we evaluate both the naïve Bayes and Robinson variants of the indicator score algorithm, optimizing thresholds for predictive performance while retaining the prior beliefs.

2.8 Ordinal Regression

Another way to build a classifier that encodes the same prior beliefs about the data as the indicator score approach is to use ordinal regression. In a manner similar to the indicator score approach, ordinal regression classifies data into one of several discrete but ordered classes. In other words, ordinal regression can encode the indicator score approach's assumption that Mention class lies between the Proto-AE and Junk classes on a spectrum. We evaluate the all-threshold loss variant of the ordinal logistic model which jointly learns the weight vector of a logistic regression model as well as the thresholds separating the classes [14].

2.9 Evaluation Metrics

Precision is the probability that a datum belongs to the positive class given that the algorithm predicted the datum belonging to the positive class. Probabilistically, if y is the true value and $\hat{y}$ is the predicted value, then:

$$\text{Precision } = p(y = 1|\hat{y} = 1)$$

Recall is the probability that the algorithm predicts a datum belonging to the positive class given that the datum belongs to the positive class:

$$\text{Recall } = p(\hat{y} = 1|y = 1)$$

The F1-score is the harmonic mean of the precision and recall:

$$F_1 = 2 * \frac{p(y = 1|\hat{y} = 1) * p(\hat{y} = 1|y = 1)}{p(y = 1|\hat{y} = 1) + p(\hat{y} = 1|y = 1)}$$

In a multiclass setting, different types of averaging can be used to compute the overall F1-score, precision, and recall. In particular, we report the micro average (which calculates the metrics globally by considering total true positives, false negatives, and false positives across classes) as well as the macro average (which is the unweighted mean of the individual metrics for each class). The micro average describes classifier performance on the data as a whole, while the macro average does not take class imbalance into account. The macro average is particularly appropriate here as two thirds of the data belong to the Junk class when the classes of highest interest are instead Proto-AE and Mention.

3 RESULTS

3.1 Spectral Embedding

We project a random 3000 item subset of the training data in count vector form to a 2-dimensional space from a 24397-dimensional space using spectral embedding with Laplacian eigenmaps. The projection is shown in Figure 1 where data points are colored by class.

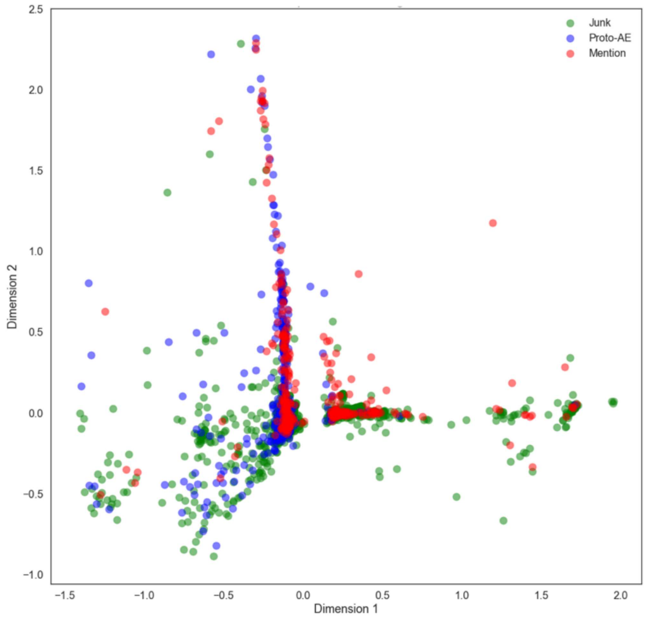


Figure 1: Spectral embedding of a training data subset.

3.2 Classification Using Indicator Scores

The model parameters and thresholds for the indicator score algorithms are learned from the training set. The performance of the naïve Bayes variant of the algorithm on the test set is reported in Table 2. The performance of the Robinson variant of the algorithm on the test set is reported in Table 3.

Table 2: The performance of the Naïve Bayes variant of the Indicator Score algorithm on the test set.

	Precision	Recall	F1-score
Junk	0.84	0.82	0.83
Proto-AE	0.59	0.79	0.67
Mention	0.12	0.07	0.09
Micro Average	0.70	0.70	0.70
Macro Average	0.51	0.56	0.53

3.3 Multinomial Naïve Bayes

The model parameters for a three-class multinomial naïve Bayes classifier are fit to the training data. The learned classifier is then run on the test data. Performance on the test data is reported in Table 4.

Table 3: The performance of the Robinson variant of the Indicator Score algorithm.

	Precision	Recall	F1-score
Junk	0.82	0.87	0.84
Proto-AE	0.73	0.71	0.72
Mention	0.25	0.19	0.22
Micro Average	0.74	0.74	0.74
Macro Average	0.60	0.59	0.59

Table 4: The performance of Multinomial Naïve Bayes.

	Precision	Recall	F1-score
Junk	**0.97**	0.69	0.81
Proto-AE	0.52	**0.93**	0.66
Mention	0.70	0.78	0.74
Micro Average	0.75	0.75	0.75
Macro Average	0.73	0.80	0.74

3.4 Logistic Regression

The model parameters for a three-class, one versus rest, L2-regularized logistic regression classifier are fit to the training data. The logistic regression classifier is then run on the test set. Classifier performance on the test set is reported in Table 5.

Table 5: The performance of Logistic Regression.

	Precision	Recall	F1-score
Junk	0.91	**0.93**	**0.92**
Proto-AE	**0.80**	0.74	**0.77**
Mention	**0.80**	**0.80**	**0.80**
Micro Average	**0.87**	**0.87**	**0.87**
Macro Average	**0.84**	**0.82**	**0.83**

3.5 Ordinal Regression

The model parameters and thresholds for an all-threshold ordinal logistic model are learned from the training data. The ordinal logistic regression classifier is then run on the test set. Classifier performance on the test set is reported in Table 6.

Table 6: The performance of Ordinal Regression.

	Precision	Recall	F1-score
Junk	0.89	0.91	0.90
Proto-AE	0.57	0.60	0.59
Mention	0.74	0.63	0.68
Micro Average	0.80	0.80	0.80
Macro Average	0.73	0.71	0.72

4 DISCUSSION

Our experiments suggest that the Robinson variant of the indicator score method performs better than the naïve Bayes variant. As such, in the following discussion, the stronger Robinson variant is used when comparing the indicator score approach to other methods. The macro average is used primarily as well in the following discussion as the majority class (Junk) is the class of least interest in a pharmacovigilance setting.

Our results show that regularized logistic regression has a better classification performance than the indicator score approach currently used by automatic pharmacovigilance systems like MedWatcher Social. In particular, on the test data, macro average precision was increased from 0.60 to 0.84, macro average recall was increased from 0.59 to 0.82, and macro average F1-score was increased from 0.59 to 0.83. In practice, these improvements translate to more efficient pharmacovigilance, better consumer safety, lower costs, and faster data availability. The performance gain can be explained by a combination of factors.

The indicator score approach implicitly assumes that the Mention class is lexically an affine combination of the Proto-AE and Junk classes. This is a strong assumption to make that results in a high bias learner when untrue. To illustrate the problem with a toy example, imagine a three-word vocabulary consisting of tokens P, M, and J. Additionally, imagine that Proto-AE documents only use token P, Mention documents only use token M, and Junk documents only use token J. Clearly, the best way to identify a Mention in this toy example is to use the M token. However, the two-class plus thresholds nature of the indicator score approach will force the algorithm to ignore token M as it is never seen in Proto-AE and Junk documents. The indicator score approach will incorrectly model the Mention class as documents that contain a certain mix of P and J tokens. This illustrates how the indicator score approach results in a high bias algorithm. The indicator score algorithm is not able to properly model the data, even in the theoretically simplest to model toy case.

It could be argued that in real world scenarios, the indicator score's affine combination assumption might be met for a subset of the vocabulary. While this is true, the indicator score approach will not be able to identify and take advantage of the subset because of the approach's lack of regularization and because the threshold optimization is performed separately from the fitting of the model's event probability parameters. In other words, for the indicator score algorithm to achieve low bias, the data would need to conform to the affine combination assumption generally across the vocabulary space. The projection of the data from the full vocabulary space to a 2-dimensional space using spectral embedding in Figure 1 strongly suggests, though does not confirm, that the affine combination assumption is not generally met. If the affine combination assumption was generally well met, we would expect the Mention class to likely broadly lie between the Proto-AE and Junk classes. Instead, we observe that the Mention and Proto-AE densities are correlated in many regions of the 2-dimensional space. This suggests that overall the vocabulary tokens used by Mention documents and Proto-AE documents are highly related, translating to a probable large number of similar class conditional event probabilities in the naïve Bayes and Robinson models, creating a need for feature importance learning and regularization not provided by the indicator score approach.

Another argument that could be made in favor of the indicator score approach is that the indicator score itself is interpretable. This argument relies on the Proto-AE class posterior computed by naïve Bayes being a well-calibrated posterior probability. Naïve Bayes is known to produce poorly calibrated class posterior probabilities [11]. The unrealistic assumption of feature independence conditioned on class distorts and pushes the produced probabilities towards the extrema of 0 and 1. Classifiers such as regularized logistic regression that do not make the assumption of feature independence conditioned on class produce better calibrated and therefore more interpretable probabilities.

Further, we note the indicator score approach with naïve Bayes and Robinsons' variant imposes a number of contradictory statistical assumptions. Intrinsically, naïve Bayes assumes that a feature x_i is conditional independent of all other features x_j given the class label y_c. The indicator approach then implicitly assumes that the middle class y_m is dependent upon the features for both other classes, since its determination is dependent on the scores for the two original classes. The indicator approach thus has internal inconsistencies in statistical modeling. The use of Robinsons' variant of using a p-value correction on what are not p-values adds another inconsistency. This would lead us to question if the indicator approach's lower performance is due to the use of naïve Bayes, combining techniques in statistically inconsistent ways, or an issue with the underlying hypothesis of a manifold / ordinal relationship in the data.

The simplest modification to make to the indicator score approach to fix the problem created by the affine combination assumption is to abandon the indicator score's use of a two-class naïve Bayes model and use a three-class naïve Bayes model instead. This eliminates the need to specify probability thresholds and reduces bias by properly modeling a three-class problem with a three-class model instead of a repurposed two-class model. A comparison of Table 2 and Table 4 confirms that abandoning the indicator score approach's use of a two-class naïve Bayes model in favor of a three-class naïve Bayes model does improve classifier performance. In particular, the macro average F1-score is increased from 0.53 to 0.74.

The projection via spectral embedding of the data suggests a second straightforward improvement to the classifier that can be made. The high density overlap between the Mention and Proto-AE classes provides evidence that it is likely that the two classes share a similar probable token vocabulary subset. The likely wide sharing of probable tokens between classes justifies a need for feature importance learning and regularization. Regularized logistic regression provides both and eliminates unrealistic feature conditional independence assumptions. Logistic regression is preferable over naïve Bayes from a model fitting perspective as well. While logistic regression and naïve Bayes fit the same probability model, they take different approaches. Logistic regression takes the discriminative approach by directly optimizing the class posterior, while naïve Bayes first fits a joint probability distribution by learning the class conditional and class prior and then second obtains the class posterior via Bayes rule. Research by Ng and Jordan has shown that given enough training data, logistic regression will outperform naïve Bayes because a discriminative approach will have a lower asymptotic error than a generative approach [10]. If naïve Bayes is performing better than logistic regression on a problem, it is likely that additional training data will be beneficial. A comparison of Table 4 and Table 5 shows that regularized logistic regression outperforms the three-class naïve Bayes algorithm. The macro average F1-score is increased from 0.74 to 0.83. Overall, regularized logistic regression when compared to the MedWatcher Social indicator score approach increases the macro average F1-score from 0.59 to 0.83 while using the same data and feature set.

Finally, we tried ordinal regression to map the original indicator score hypothesis onto a more formally developed framework. Table 6 shows that ordinal logistic regression performs better than both the naïve Bayes and Robinson variants of the indicator score approach, but also that ordinal logistic regression performs worse on all metrics when compared to logistic regression. Logistic regression outperforming ordinal logistic regression provides further evidence that the original indicator score hypothesis of ordered classes does not hold well in this scenario. The improvement in performance seen when switching from the indicator score approach to ordinal logistic regression indicates that the original assumption is potentially partially validated, and a partial ordinal relationship could exist within the data. However, the ordinal relationship does not accurately model the entire underlying data distribution. Given sufficient data, a model without the indicator/ordinal assumption appears to have the best performance.

We note that logistic regression is a simple, standard machine learning tool and that many more complex classification algorithms exist. Future avenues of research include exploring alternative feature extraction procedures, models, and inference techniques for adverse event identification with this data. A standardized comparison and characterization of automatic pharmacovigilance systems using common training and testing data would also be a beneficial avenue of future research. As pointed out by Sarker et al., it is unfortunately difficult currently to rigorously compare the performance of different social listening driven pharmacovigilance systems because various data sources, data sizes, and gold standards were used in the different research studies [18].

5 CONCLUSION

This study examined the MedWatcher Social indicator score algorithm for automatic pharmacovigilance using social media data. An analysis of the indicator score approach's modeling assumptions suggests that performance can be improved by switching to a regularized, three-class logistic regression. Careful consideration of modeling assumptions and regular reevaluation are critical to the design and proper use of algorithms for computational pharmacovigilance and computational epidemiology.

More broadly, the pharmacovigilance community has developed some of its own tools, and some of these tools have existing counterparts in the machine learning space. In many fields, the judicious application of machine learning often outperforms community built tools. We have demonstrated an example of when this is the case.

ACKNOWLEDGEMENTS

Special thanks to Carrie Pierce and Clark Freifeld for interesting conversations and insights.

REFERENCES

[1] Belkin, M. and Niyogi, P., 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. In Advances in neural information processing systems (pp. 585-591).

[2] Belkin, M. and Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, 15(6), pp.1373-1396.

[3] U.S. Food and Drug Administration., 2017. FDA Adverse Event Reporting System (FAERS): Latest Quarterly Data Files. Available at: https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm082193.htm [Accessed 26 Dec. 2017].

[4] Fisher, R.A., 1925. Statistical methods for research workers. Genesis Publishing Pvt Ltd.

[5] Freifeld, C.C., Brownstein, J.S., Menone, C.M., Bao, W., Filice, R., Kass-Hout, T. and Dasgupta, N., 2014. Digital drug safety surveillance: monitoring pharmaceutical products in twitter. Drug safety, 37(5), pp.343-350.

[6] Freifeld, C.C., 2014. Digital pharmacovigilance: The medwatcher system for monitoring adverse events through automated processing of internet social media and crowdsourcing (Doctoral dissertation, Boston University).

[7] Levinson, D.R., 2010. Adverse events in hospitals: national incidence among Medicare beneficiaries. Department of Health and Human Services Office of the Inspector General.

[8] Levinson, D.R., 2012. Hospital incident reporting systems do not capture most patient harm. Washington DC: Office of the Inspector General.

[9] Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.

[10] Ng, A.Y. and Jordan, M.I., 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Advances in neural information processing systems (pp. 841-848).

[11] Niculescu-Mizil, A. and Caruana, R., 2005, August. Predicting good probabilities with supervised learning. In Proceedings of the 22nd international conference on Machine learning (pp. 625-632). ACM.

[12] Pierce, C.E., Bouri, K., Pamer, C., Proestel, S., Rodriguez, H.W., Van Le, H., Freifeld, C.C., Brownstein, J.S., Walderhaug, M., Edwards, I.R. and Dasgupta, N., 2017. Evaluation of Facebook and Twitter monitoring to detect safety signals for medical products: an analysis of recent FDA safety alerts. Drug safety, 40(4), pp.317-331.

[13] Powell, G.E., Seifert, H.A., Reblin, T., Burstein, P.J., Blowers, J., Menius, J.A., Painter, J.L., Thomas, M., Pierce, C.E., Rodriguez, H.W. and Brownstein, J.S., 2016. Social media listening for routine post-marketing safety surveillance. Drug safety, 39(5), pp.443-454.

[14] Rennie, J.D. and Srebro, N., 2005, July. Loss functions for preference levels: Regression with discrete ordered labels. In Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling (pp. 180-186). Kluwer Norwell, MA.

[15] Robinson, G., 2003. A statistical approach to the spam problem. Linux journal, 2003(107), p.3.

[16] Santillana, M., Zhang, D.W., Althouse, B.M. and Ayers, J.W., 2014. What can digital disease detection learn from (an external revision to) Google Flu Trends?. American journal of preventive medicine, 47(3), pp.341-347.

[17] Santillana, M., Nguyen, A.T., Dredze, M., Paul, M.J., Nsoesie, E.O. and Brownstein, J.S., 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS computational biology, 11(10), p.e1004513.

[18] Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T. and Gonzalez, G., 2015. Utilizing social media data for pharmacovigilance: A review. Journal of biomedical informatics, 54, pp.202-212.

[19] World Health Organization, 2002. The importance of pharmacovigilance.

Machine learning in healthcare - a system's perspective

[Position paper]

Awais Ashfaq
Center of Applied Intelligent Systems Research,
Halmstad University. Sweden
Halland Hospital, Region Halland. Sweden
awais.ashfaq@hh.se

Slawomir Nowaczyk
Center of Applied Intelligent Systems Research,
Halmstad University. Sweden
slawomir.nowaczyk@hh.se

ABSTRACT

A consequence of the fragmented and siloed healthcare landscape is that patient care (and data) is split along multitude of different facilities and computer systems and enabling interoperability between these systems is hard. The lack interoperability not only hinders continuity of care and burdens providers, but also hinders effective application of Machine Learning (ML) algorithms. Thus, most current ML algorithms, designed to understand patient care and facilitate clinical decision-support, are trained on limited datasets. This approach is analogous to the Newtonian paradigm of *Reductionism* in which a system is broken down into elementary components and a description of the whole is formed by understanding those components individually. A key limitation of the reductionist approach is that it ignores the component-component interactions and dynamics within the system which are often of prime significance in understanding the overall behaviour of complex adaptive systems (CAS). Healthcare is a CAS.

Though the application of ML on health data have shown incremental improvements for clinical decision support, ML has a much a broader potential to restructure care delivery as a whole and maximize care value. However, this ML potential remains largely untapped: primarily due to functional limitations of Electronic Health Records (EHR) and the inability to see the healthcare system as a whole. This viewpoint (i) articulates the healthcare as a complex system which has a biological and an organizational perspective, (ii) motivates with examples, the need of a system's approach when addressing healthcare challenges via ML and, (iii) emphasizes to unleash EHR functionality - while duly respecting all ethical and legal concerns - to reap full benefits of ML.

CCS Concepts

•Computing methodologies → Machine learning; Systems theory;

epiDAMIK: Epidemiology meets Data Mining and Knowledge discovery
August 5, 2019, Anchorage Alaska, USA
© 2019 ACM. ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

Keywords

Machine learning; Healthcare complexity; System's thinking; Electronic health records

Introduction

System's thinking is a holistic approach to understand Complex Adaptive Systems (CAS) by not only focusing on individual components of the system but also the interconnections between those components [1]. A CAS is (i) a collection of several individual agents, (ii) without centralised co-ordination, (iii) with freedom to act in ways that are not always totally predictable and, (iv) whose actions are interconnected so that one agent's actions change the context for other agents resulting in novel characteristics exhibited by the system as a whole. Put differently, the overall output of a CAS is not equal to the sum of outputs of all its subsystems. Examples include, but not limited to, the global climate, the financial market and the healthcare system.

Since the beginning of the 21st century, healthcare has been widely framed and studied as CAS on many levels such as diseases, patients, epidemiology, care-teams, hierarchy, practices and education to unravel the complexity and improve decision-making [2–5]. The complexity triggers in healthcare can be broadly described from a biological (*intra-human* e.g. the immune system) and organizational (*inter-human* e.g. multidisciplinary care teams) perspective.

From a biological perspective, an important source of complexity is the uncertainty in medical knowledge: explained by the complex human biology which is subject to nearly constant change within physiological pathways due to a complex series of gene/environment interactions [6]. Concurrently, International Classification of Diseases (ICD-10) specified over 68,000 diagnoses and the list keeps growing as we await ICD-11. In order to cure or alleviate patient sufferings, clinicians practice thousands of drugs and therapies. As a result, the gap between the rapidly advancing medical knowledge base and the application of that knowledge to escalating population size continues to widen [7]. Since clinical decisions deal with human life, we want to be as certain as possible about practice and desired clinical outcomes. Put differently, unravelling biological complexity aims at reducing the uncertainty in clinical decisions.

From an organizational perspective, an important source of complexity is healthcare structure, which - unlike an engineered complex system - is a system that emerged over time with independent actors (patients, care-providers, technologists, tax-payers; and hospitals, clinics, laboratories, government etc.) involved [8]. The actors often have distinct leaderships, budgets, goals, regulations and tools of operation. Controlling the output of such a socio-technical complex system is - if not impossible - very challenging because of the high degree of inter-relatedness among the roles of actors which renders the overall system output not equal to the sum of outputs of all the sub-systems. Put differently, unravelling organizational complexity aims at identifying problem areas in healthcare where interventions can have significant impact on the overall system output.

Healthcare complexity and Machine Learning

The application of ML in healthcare is widely anticipated as a key step towards improving care quality and curbing care costs [9]. A boon to this anticipation is the widespread adoption of Electronic Health Records (EHRs) in the health system. In Sweden, EHRs were introduced in the 1990s and by 2010, 97% of hospitals, and 100% of primary care doctors used them for their practice [10]. EHRs are real-time digital patient-centred records that allow secure access to authorized care-providers across multiple healthcare centres when required. This digitization in healthcare generates unprecedented amounts of clinical data, which when coupled with modern ML tools provides an opportunity to expand the evidence base of medicine and facilitate decision process.

For instance, inpatient crowding (or high levels of hospital occupancy) are often associated with reduced quality of care and access burden on care-providers [11]. The benefits of discharge are both monetary: hospital stays are expensive, bed availability is increased; and non-monetary: patients are less prone to medical complications and can spend more time with family and return to work. However the benefits depend on patient outcome (prognosis), so discharging a patient is worthy only if he or she does not need to be re-admitted in near future, which makes it a prediction problem. Put differently, the ML predictive challenge is: which patients are likely to be readmitted using data available at the point of discharge.

Similarly, for instance, long waiting times for patients to access the next step in the care-process has long been a cause of dissatisfaction for patients and care-providers [12]. The benefits of reducing waiting times are several: better value for patient's time, less work stress among care-providers, improved care quality and quick monetary reimbursements. However to reduce waiting times, we want to predict demand - or patient path - in order to allocate resources (staff, drugs, equipment etc.) efficiently. Put differently, the ML predictive challenge is: what path will the patient follow in the healthcare system using data available at the time of entry.

In the current era, where algorithms can beat GO champions [13] or drive a man to the hospital [14], it is tempting to believe that the aforementioned ML challenges are not far-fetched. However, it is worth remembering that the utility of ML tools largely hinges on underlying data and in most situations, access to complete healthcare records is extremely challenging. Though there are inklings of AI in medicine, the necessary resources - data - are still lacking. The Dataset Information Resource[1] (DIR) [15] describes over 12 commonly used EHR datasets for research projects, of which only 2 are publicly available [16, 17]. The more comprehensive datasets with 100,000+ subjects and integrating healthcare information from diverse care points are proprietary and often require approvals from one or more advisory committees along with access charges [18–22].

Advancements in medical knowledge and guidelines have progressed a paradigm shift in healthcare: from care in a single unit to care across multiple units with varying but specialized expertise. It is often referred as fragmentation in medicine. Thus, patient care (and of course data) is split along multitude of different facilities and computer systems and integration of this information into a single system faces numerous challenges, primarily from an organizational perspective [23]. These include privacy and security concerns, lack of acceptable standardized data formats, use of proprietary technologies by disparate vendors, costly interface fees and more. As a result, care-providers are impeded from accessing complete datasets and thus unable to understand all aspects of the patient health journey.

Just as humans are better equipped to understand the world when given complete facts, so too are algorithms. As a consequence of the fragmented and siloed landscape of healthcare, the potential of ML algorithms to understand care patterns remains largely untapped, both due to *unavailability* and *inaccessibility* of necessary data. For instance, one barrier to prediction studies (such as in the readmission example) is that patient information after the course of treatment - the true outcome data - is not always available in EHR. Though, patient reported outcome and experience measures are being developed and validated, their integration into EHRs is scarce [24]. Data from health devices at homes – that monitor patient's health beyond the hospital radar - are also an important source of information, yet their integration to existing EHRs is a challenge [25].

Similarly, one hurdle to predict patient flows is the lack of interoperability between EHRs in different care chains. In healthcare, we consider care fragments (primary, secondary, emergency etc.) as independent bodies; however, they constitute a single body from a patient's perspective who travels through them during the care process. Thus addressing the ML predictive challenges would require accessing and merging data from all levels of the care chain to have a holistic (complete) approach to healthcare delivery. The urge of a holistic approach is highly emphasized today; else - given the huge number of ML applications in healthcare - we might soon face another wave of interoperability challenges. Only this time, it will be between ML prediction models for different care fragments.

Text focus and context demise

The process of training most current ML algorithms on limited and often different health datasets, to understand patient care and facilitate decision-support is, to a large extent, analogous to the Newtonian paradigm of a *clockwork universe* that aims to understand a system by breaking it down into elementary components and understanding those components to form a description of the whole. This mode of scientific inquiry is often referred as Reductionism (as opposed to Holism) following the belief that any system can be explained by analysing the most basic components of the system. Reductionism has guided scientific reasoning for centuries with great success such as the development of the cell theory or formulation of the periodic table etc. that are fundamentals of explaining a wide variety of things that we encounter in our lives. Despite being a very powerful ap-

[1] https://cci-hit.uncc.edu/dir/index.php/Welcome_to_DIR

proach in explaining individual components, Reductionism struggles to understand the new emergent behaviour that results due to non-linear interactions between the individual components. In the context of ML on limited EHR data, one might consider a ML model trained to predict hospital admission given Emergency Department (ED) data. The developed model might exhibit strong discrimination ability and rightfully recommend transfer of a sicker patient from ED to the hospital. However, the utility of this decision largely hinges on the availability of resources in the hospital (beds, care-providers) at the time of admission and transferring a patient without assuring resource availability in the hospital might deteriorate the health state of an already sick patient.

Conclusion

EHRs are, in sum, valuable resources for clinical and organizational decision-making in healthcare. However, in order to reap full benefits of ML in healthcare, we need to realize the limitations of existing datasets and appreciate a system's approach to model the complex healthcare landscape. In the context of healthcare data, a system's perspective would mean a comprehensive data resource covering complete clinical, operational and financial information of care processes at individual, organizational and population levels. The Institute of Medicine also recommends including social and behavior measures into patient EHRs (stress, isolation levels, physical activity, geocoding etc.) to better characterize the diverse range of factors that drive complex diseases and influence individual and population health [26]. The challenge of interoperability among different EHR systems is, albeit hard but, not insurmountable [19, 27]. Simultaneously, the rejiggering of EHRs would demand novel research challenges in fields of data security and differential privacy to create scalable and secure data warehouses to store sensitive medical information and facilitate responsible access to researchers when required. Dual-purposing of EHRs also demands reorientation of ethical and legal rules because in addition to medical professionals, EHRs are being accessed by data scientists and administrators.

References

[1] Stephen Wolfram. *A new kind of science*, volume 5. Wolfram media Champaign, IL, 2002.

[2] Olaf Dammann, Phillip Gray, Pierre Gressens, Olaf Wolkenhauer, and Alan Leviton. Systems epidemiology: what's in a name? *Online journal of public health informatics*, 6(3), 2014.

[3] Paul E Plsek and Trisha Greenhalgh. The challenge of complexity in health care. *Bmj*, 323(7313):625–628, 2001.

[4] Paul E Plsek and Tim Wilson. Complexity, leadership, and management in healthcare organisations. *Bmj*, 323 (7315):746–749, 2001.

[5] David JD Earn, Pejman Rohani, Benjamin M Bolker, and Bryan T Grenfell. A simple model for complex dynamical transitions in epidemics. *science*, 287(5453): 667–670, 2000.

[6] Peter Densen. Challenges and opportunities facing medical education. *Transactions of the American Clinical and Climatological Association*, 122:48, 2011.

[7] World Health Organization. *World report on ageing and health*. World Health Organization, 2015.

[8] William B Rouse. Health care as a complex adaptive system: implications for design and management. *Bridge-Washington-National Academy of Engineering-*, 38(1):17, 2008.

[9] Geoffrey Hinton. Deep learning—a technology with the potential to transform health care. *Jama*, 320(11):1101–1102, 2018.

[10] D Adamski. Overview of the national laws on electronic health records in the eu member states. national report for poland, 2014.

[11] Joel S Weissman, Jeffrey M Rothschild, Eran Bendavid, Peter Sprivulis, E Francis Cook, R Scott Evans, Yevgenia Kaganova, Melissa Bender, JoAnn David-Kasdan, Peter Haug, et al. Hospital workload and adverse events. *Medical care*, 45(5):448–455, 2007.

[12] Bernd Rechel, Martin McKee, Marion Haas, Gregory P Marchildon, Frederic Bousquet, Miriam Blümel, Alexander Geissler, Ewout van Ginneken, Toni Ashton, Ingrid Sperre Saunes, et al. Public reporting on quality, waiting times and patient experience in 11 high-income countries. *Health Policy*, 120(4):377–383, 2016.

[13] Elizabeth Gibney. Google ai algorithm masters ancient game of go. *Nature News*, 529(7587):445, 2016.

[14] Alfonso Reed, Richmond Sakie, and Demonyae Smith. Umb cure 2018: Self-driving cars. 2018.

[15] Jingyi Shi, Mingna Zheng, Lixia Yao, and Yaorong Ge. Developing a healthcare dataset information resource (dir) based on semantic web. *BMC medical genomics*, 11(5):102, 2018.

[16] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[17] NHANES. National health and nutrition examination survey. 2019.

[18] Truven Health. Marketscan research data. 2019.

[19] Achim Wolf, Daniel Dedman, Jennifer Campbell, Helen Booth, Darren Lunn, Jennifer Chapman, and Puja Myles. Data resource profile: Clinical practice research datalink (cprd) aurum. *International journal of epidemiology*, 2019.

[20] Vidhya Gunaseelan, Brooke Kenney, Jay Soong-Jin Lee, and Hsou Mei Hu. Databases for surgical health services research: Clinformatics data mart. *Surgery*, 165(4):669–671, 2019.

[21] THIN. The health imrovement network. 2019.

[22] PHD. Premier healthcare databse. 2019.

[23] Afeezat Olajumoke Oyeyemi and Philip Scott. Interoperability in health and social care: organizational issues are the biggest challenge. *Journal of innovation in health informatics*, 25(3):196–198, 2018.

[24] Heather Taffet Gold, Raj J Karia, Alissa Link, Rachel Lebwohl, Joseph D Zuckerman, Thomas J Errico, James D Slover, Aaron J Buckland, Devin M Mann, and Michael N Cantor. Implementation and early adaptation of patient-reported outcome measures into an electronic health record: A technical report. *Health informatics journal*, page 1460458218813710, 2018.

[25] Nicholas Genes, Samantha Violante, Christine Cetrangol, Linda Rogers, Eric E Schadt, and Yu-Feng Yvonne Chan. From smartphone to ehr: a case report on integrating patient-generated health data. *npj Digital Medicine*, 1(1):23, 2018.

[26] Institute of Medicine (US). Committee on the Recommended Social, Behavioral Domains, and Measures for Electronic Health Records. *Capturing social and behavioral domains and measures in electronic health records: phase 2.* National Academies Press, 2014.

[27] Awais Ashfaq and Stefan Lönn et al. Data resource profile: Regional healthcare information platform in sweden (pre-print submitted). *International journal of epidemiology*, 2019.

Improving Outbreak Detection with Stacking of Statistical Surveillance Methods

Moritz Kulessa
Knowledge Engineering Group
Technische Universität Darmstadt
Germany
mkulessa@ke.tu-darmstadt.de

Eneldo Loza Mencía
Knowledge Engineering Group
Technische Universität Darmstadt
Germany
eneldo@ke.tu-darmstadt.de

Johannes Fürnkranz
Knowledge Engineering Group
Technische Universität Darmstadt
Germany
juffi@ke.tu-darmstadt.de

ABSTRACT

Epidemiologists use a variety of statistical algorithms for the early detection of outbreaks. The practical usefulness of such methods highly depends on the trade-off between the detection rate of outbreaks and the chances of raising a false alarm. Recent research has shown that the use of machine learning for the fusion of multiple statistical algorithms improves outbreak detection. Instead of relying only on the binary output (*alarm* or *no alarm*) of the statistical algorithms, we propose to make use of their p-values for training a fusion classifier. In addition, we also show that adding additional features and adapting the labeling of an epidemic period may further improve performance. For comparison and evaluation, a new measure is introduced which captures the performance of an outbreak detection method with respect to a low rate of false alarms more precisely than previous works. Our results on synthetic data show that it is challenging to improve the performance with a trainable fusion method based on machine learning. In particular, the use of a fusion classifier that is only based on binary outputs of the statistical surveillance methods can make the overall performance worse than directly using the underlying algorithms. However, the use of p-values and additional information for the learning is promising, enabling to identify more valuable patterns to detect outbreaks.

ACM Reference Format:
Moritz Kulessa, Eneldo Loza Mencía, and Johannes Fürnkranz. 2019. Improving Outbreak Detection with Stacking of Statistical Surveillance Methods. In *Proceedings of epiDAMINK 2019: Epidemiology meets Data Mining and Knowledge discovery, Workshop held in conjuction with ACM SIGKDD 2019 (epiDAMINK '19).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The early detection of infectious disease outbreaks is of great significance for public health. The spread of such outbreaks could be diminished tremendously by applying control measures as early as possible, which indeed can save lives and reduce suffering [19]. For that purpose, statistical algorithms have been developed to automate and improve outbreak detection. Such methods raise alarms in the case that an unusually high number of infections is detected which results in a further investigation by an epidemiologist [10]. Ideally, such algorithms are completely automated while still being able to be applied on a wide spectrum of different infections and syndromes [20]. However, if not chosen wisely or configured properly, they may also raise many false alarms which can overwhelm the epidemiologist. In particular for large surveillance systems, where many time series for different diseases and different locations are monitored simultaneously, the false alarm rate is a major concern and therefore highly determines the practical usefulness of an outbreak detection method [23]. However, regulating the false alarm rate usually has an impact on the ability to detect outbreaks. To find a good trade-off between those measures is one of the major challenges in outbreak detection [1, 19].

Traditional outbreak detection methods rely on historic data to fit a parametric distribution which is then used to check the statistical significance of the current observation. Choosing the significance level for the statistical method beforehand makes the evaluation difficult. In line with Kleinman and Abrams [15], we propose a method which uses the p-values of the statistical methods in order to evaluate their performance. In particular, we propose a variant of Receiver Operating Characteristic (ROC) curves, which shows the false alarm rate on the x-axis and the detection rate—in contrast to the true positive rate—on the y-axis. By using the area under the *partial* ROC curve [17], we are able to obtain a measure for the performance of an algorithm satisfying a particular constraint on the false alarm rate (e.g. less than 1% false alarms). This criterion serves as the main measure for our evaluations and enables to analyze the trade-off between the false alarm rate and the detection rate of outbreak detection methods precisely.

Prior work on outbreak detection mainly focuses on forecasting the number of infections for a disease (e.g. [3, 4]). However, only little research has been devoted to use supervised machine learning (ML) techniques for improving algorithms, which can raise alarms. Jafarpour et al. [11] used *Baysian networks* to identify the determinants for detection performance to find appropriate algorithm configurations for outbreak detection methods. Furthermore, classification algorithms and voting schemes have been used for the fusion of outbreak detection methods on univariate time series [12, 24] as well as on multi-stream time series [2, 16, 18]. However, the examined approaches only rely on the binary output (*alarm* or *no alarm*) of the underlying statistical methods for the fusion which limits the information about a particular observation.

epiDAMINK '19, August 05, 2019, Anchorage, Alaska - USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Prior research in the area of ML has shown that more precise information of the underlying models improves the overall performance of the fusion [25]. Therefore, we propose an approach for the fusion of outbreak detection methods which uses the p-values of the underlying statistical methods. Moreover, one can also incorporate different information for the outbreak detection (e.g., weather data, holidays, statistics about the data, ...) by just augmenting the data with additional attributes. As a first step, we put our focus on improving the performance of outbreak detection methods using an univariate time series as the only source of information. Furthermore, the way outbreaks are labeled in the data also has a major influence on the learnability of outbreak detectors. Thus, we propose adaptions for the labeling of outbreaks in order to maximize the detection rate of ML algorithms.

2 STATISTICAL ALGORITHMS FOR SYNDROMIC SURVEILLANCE

The key idea of our approach is to learn to combine predictions of commonly used statistical outbreak detection methods with a trainable ML algorithm. Thus, we first need to generate a series of aligned prediction vectors, each consisting of one entry for each method. This sequence can then be used for training the ML model.

Let us denote with $C = (c_0, c_1, \ldots, c_n) \in \mathbb{N}^n$ the time series of infection counts for a particular disease. Many methods rely on a sliding window approach which uses the previous m counts as reference values for fitting a particular parametric distribution. Therefore, the mean $\mu(t)$ and the variance $\sigma^2(t)$ can be computed over these m reference values as follows:

$$\mu(t) = \frac{1}{m}\sum_{i=1}^{m} c_{t-i} \qquad \sigma^2(t) = \frac{1}{m}\sum_{i=1}^{m}(c_{t-i} - \mu)^2$$

On the fitted distributions, a statistical significance test is performed in order to identify suspicious spikes. For the purpose of outbreak detection, we rely on one tailed-tests for the statistical algorithms in order to only capture the observation of unusual high number of infections. For a particular observed count c_t and a fitted distribution $p(x)$, the p-value is computed as the probability $\int_{c_t}^{\infty} p(x)dx$ of observing c_t or higher counts. Hence, small p-values represent uncommonly high counts of c_t. The sensitivity of raising an alarm is regulated by the significance level α and if the p-value is inferior to the threshold α an alarm is raised.

We have chosen to base our work on the following methods which are all implemented in the R package *surveillance* [22]:

EARS C1 and **EARS C2** are variants of the *Early Aberration Reporting System* [7, 9] which rely on the assumption of a Gaussian distribution. The difference between C2 and C1 lies in the added gap of two time points between the reference values and the current observed count c_t, so that the distribution of c_t are assumed as in the following:

$$c_t \overset{\text{C1}}{\sim} N(\mu(t), \sigma^2(t)) \qquad c_t \overset{\text{C2}}{\sim} N(\mu(t-2),\ \sigma^2(t-2))$$

EARS C3 combines the result of the C2 method over a period of three previous observations. For convenience of notation, the incidence counts c_t for the C3 method are transformed according to the statistics so that it fits to the normal distribution.

$$\left[\frac{c_t - \mu(t-2)}{\sqrt{\sigma^2(t-2)}} - \sum_{i=1}^{2}\max(0, \frac{c_{t-i} - \mu(t-2-i)}{\sqrt{\sigma^2(t-2-i)}} - 1)\right] \overset{\text{C3}}{\sim} N(0,1)$$

Despite the inaccurate assumption of the Gaussian distribution for low counts, the EARS variants are often included in comparative studies due to its simplicity and still serves as competitive baseline [1, 7, 8].

Bayes method. In contrast to the family of C-algorithms, the Bayes algorithm relies on the assumption of a negative binomial distribution:

$$c_t \overset{\text{Bayes}}{\sim} NB(m \cdot \mu(t) + \frac{1}{2},\ \frac{m}{m+1})$$

RKI method. Since the Gaussian distribution is not suitable for count data with a low mean, the RKI algorithm, as implemented by Salmon et al. [22], assumes a Poisson distribution:

$$c_t \overset{\text{RKI}}{\sim} \begin{cases} Poisson(\lfloor \mu(t) \rfloor + 1), & \text{if } \mu(t) \le 20 \\ N(\mu(t),\ \sigma^2(t)), & \text{otherwise} \end{cases}$$

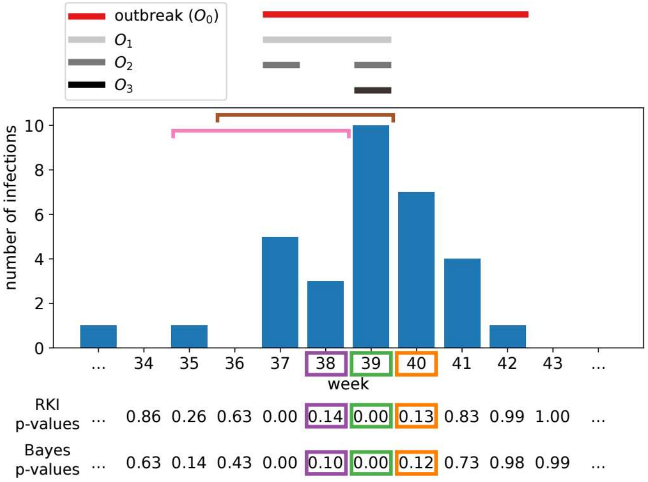
They all have in common that they require comparably little historic data on their own, which allows us to train the ML method on longer sequences. Moreover, such methods are universally applicable and serve as drop-in approaches for surveillance systems since they only rely on the detection of a local increase in incidents without the need to capture effects like seasonality and trend.

3 FUSION METHODS

The combination of information from several sources in order to obtain a unified picture is known as *fusion* [14]. *Classifier fusion* is a special case which combines the outputs of multiple classifiers in order to improve classification performance. In our context, the statistical algorithms for syndromic surveillance can be seen as classifiers, each classifying the current observation into the classes *alarm* or *no alarm*. A straight-forward way for combining the predictions of multiple outbreak detection methods is to simply vote and follow the majority prediction. A more sophisticated approach consists of training a classifier that uses the predictions of the detection methods as input, and is trained on the desired output, a technique that is known in ML as *stacking* [26].

Recent work in the area of outbreak detection and fusion has focused on fusing the information obtained by simultaneously monitoring multiple time series for a particular disease. Lau et al. [16] have shown that the performance of statistical algorithms can already be improved by combining them with simple voting schemes. Mnatsakanyan et al. [18] could further improve the performance using Bayesian networks and including further information about the patients (e.g., age) as additional attributes. Moreover, Burkom et al. [2] have used a hierarchy of Bayesian networks in order to incorporate additional information about health surveillance data and environmental sensors. However, all of these fusion methods aim to capture the degree of dependence between the monitored time series relying on spatial correlations.

Only little research has been devoted to improving the performance of statistical algorithms on univariate time series. In particular, Texier et al. [24] have used the ML technique *hierarchical*



index	augmented features			p-values		
week		prev. p-values				target
t	mean$_t$	RKI$_{t-1}$	Bayes$_{t-1}$	RKI$_t$	Bayes$_t$	outbreak$_t$
...	...	...	...	...	...	...
34	1.00	0.59	0.63	0.86	0.63	no
35	0.50	0.86	0.63	0.26	0.14	no
36	0.50	0.26	0.14	0.63	0.43	no
37	0.50	0.63	0.43	0.00	0.00	yes
38	1.50	0.00	0.00	0.14	0.10	yes
39	2.25	0.14	0.10	0.00	0.00	yes
40	4.50	0.00	0.00	0.13	0.12	yes
41	6.25	0.13	0.12	0.83	0.73	yes
42	6.00	0.83	0.73	0.99	0.98	yes
43	5.50	0.99	0.98	1.00	0.99	no
...	...	...	...	...	...	...

Figure 1: Example for the creation of training data for the learning algorithm including the statistical algorithms Bayes and RKI with a window size of one ($w = 1$) and the mean over the previous four counts ($m = 4$) as features. On the left hand side, the time series for a particular disease is visualized at the center representing the number of cases of infections over time. The computed p-values of the statistical algorithms (underneath) and the label indicating an outbreak for each observation (above) are placed at the respective time index t. Using this information the data instances can be created as shown on the right: Each particular time point is represented by one training instance, labeled according to the original targets O_0.

mixture of experts [13] to combine the output of the methods from EARS. However, the authors note that all algorithms rely on the assumption of a Gaussian distribution, which limits their diversity. In contrast, Jafarpour et al. [12] have used a variety of classification algorithms (*logistic regression*, *CART* and *Baysian Networks*) for the fusion of outbreak detection methods. As underlying statistical algorithms they have used the Cumulative Sum (CUSUM), two Exponential Weighted Moving Average algorithms, the EARS methods (C1,C2,C3) and the Farrington algorithm [19]. In general, the results of Texier et al. [24] and Jafarpour et al. [12] indicate that ML improves the ability to detect outbreaks while simple voting schemes (e.g. weighted voting and majority vote) did not perform well. Moreover, the algorithms have not been evaluated with respect to data which include seasonality and trend.

4 FUSION WITH AUGMENTED STACKING

In this work, we show that the availability of additional information can further improve the performance of the fusion classifier. Therefore, we first propose to use p-values of the statistical methods for the fusion in order to include information about the certainty of an alarm, and then show how to add additional external information to the learning process of the ML algorithm. Finally, we investigate different variants for labeling outbreaks.

4.1 Fusion with p-values

Given base estimators $g_1(x), \dots, g_K(x)$, a *fusion combiner* is a function $h(g_1(x), \dots, g_K(x))$ that combines the predictions of the base functions. In the simple case of binary voting, i.e., $g_i(x) \in \{0, 1\}$, the combiner $h(x) = \frac{1}{K}\sum_i g_i(x)$ with a threshold of 0.5 would model the majority rule. In *stacking* the function $h : X^K \longrightarrow O$ is learned by training a machine learning classifier on a set of previous observations $(g_1(x_1), \dots, g_K(x_1)), \dots, (g_1(x_n), \dots, g_K(x_n))$ −derived from applying g_i on x_t− with associated targets $o_1, \dots, o_n \in O$. We refer to this as the training set in contrast to the evaluation set, which contains new, unseen observations. In outbreak detection, the instances x_t correspond to the points in the time series C of infection counts c_t and $o_t \in \{0, 1\}$ denotes the labelling of a time point as belonging to an outbreak (1) or not (0).

Previous approaches [12, 24] used the binary alarms ({0,1}) of base outbreak detectors. In this work instead, we propose to base our stacking model on the p-values, i.e., $g_i(x) \in [0, 1]$, provided by the underlying statistical approaches (cf. Sec. 2). In fact, the p-values can directly be seen as the certainty of currently observing an outbreak, enabling the learning algorithm to make use of the base estimations in a much more fine grained way. This information is otherwise lost when using binary alarms, which are indeed obtained by just applying a fixed threshold on the computed p-values. In addition to the circumvented difficulty of tuning such threshold, previous studies on stacking have shown empirically that using the raw predictions can improve over the discretized option [25].

Figure 1 visualizes an example on how the data for the learning algorithm is created by using the p-values of the statistical algorithms Bayes and RKI. The columns RKI$_t$ and Bayes$_t$ represent the computed p-values for the current observation while the other columns (mean$_t$, RKI$_{t-1}$ and Bayes$_{t-1}$) represent additional information explained in the following section.

4.2 Additional Features

The use of a trainable fusion method allows us to include additional information which can help to decide whether a given alarm should be raised or not. As additional features, we propose to include the *mean* of the counts over the last m time points (the same number of time points as used by the statistical methods), which can give us evidence about the reliability of a particular outcome. For example, the assumption of a Gaussian distribution for a low mean of count data (≤ 20) is known to be imprecise. Therefore, a learning algorithm might induce in this scenario that the p-values of the statistical methods C1, C2 and C3 may not be trustworthy. Moreover, under the assumption that a time series is stationary an unusual high mean can also be a good indicator to detect an outbreak, especially in the case that an outbreak arises slowly over time. The column mean_t in Figure 1 illustrates how the mean over the last four observed counts ($m = 4$) is added as an additional feature.

Finally, we also include the output of the statistical methods for previous time points in a window of a user-defined size w as additional features. For the example in Figure 1, we have used a window size of one ($w = 1$) which includes the previous output of both statistical algorithms.

4.3 Modelling the Output Labels for Learning

A major challenge for ML algorithms is that the duration of an outbreak period is not clearly defined [23]. A simple strategy—which we refer to as O_0—is to label all time points positive as long as cases for the particular epidemic are reported (e.g. time points prior to the peak of an outbreak and a few time points after the peak). In this case, the goal of the learning algorithm is to predict most time points in an ongoing epidemic as positive, regardless of their time stamp. Indeed, our early results indicate that the predictor learns to recognize the fading-out of an outbreak (e.g. weeks 40 to 42 in Figure 1). This is due to the fact that the peak of the outbreak is included in the reference values which results in a considerably high mean $\mu(t)$ for the significance test. Because of this, unusually high p-values are generated for the counts after the peak, which provide sufficient evidence for the stacking algorithm to raise an alarm. However, this also increases the number of false alarms as the ML approach learns to raise alarms when the count is decreasing outside an epidemic period.

To avoid this, we propose three adaptations of O_0: O_1 labels all time points until the peak (the point with maximum number of counts during the period) as positive. O_2 instead skips the time points whose count is decreasing compared to the immediate previous count (i.e., it labels all increasing counts until reaching the peak). Finally, O_3 labels only the peak of the outbreak as positive. Figure 1 visualizes an example outbreak with the corresponding different options to label the epidemic period on the top-left.

5 EVALUATION MEASURES

Instead of manually adjusting the α parameter of the statistical methods and examining the results individually, which is mostly done in previous works, we propose to evaluate the p-value as it is done by Kleinman and Abrams [15]. In particular, the p-value can be interpreted as a score, which sorts examples according to their degree to which they indicate an alarm. This allows us to analyze an algorithm with ROC curves [5]. A ROC curve can be used to examine the trade-off between the *true positive rate* (i.e., the probability of raising an alarm in case of an actual outbreak) and the *false alarm rate* (i.e., the probability of falsely raising an alarm when no outbreak is ongoing). In order to only focus on high specificity results (e.g., with a false alarm rate below 1%), which is of major importance for many medical applications, we only consider *partial ROC curves*. By using the partial area under the ROC curve as proposed in [17], we obtain a simple measure to evaluate the performance of an algorithm, satisfying particular constraint on the false alarm rate. We refer to this measure as $pAUC_e$ where the parameter e defines the maximum allowed false alarm rate to be considered. It is computed as

$$pAUC_e = \frac{\int_0^e ROC(f)\,df}{e}$$

where $ROC(f)$ denotes the true positive rate given a false alarm rate of f. However, alarms raised in cases when the epidemic has already been detected are typically not very decisive and informative anymore. To incorporate this, we consider the *detection rate*, which represents the proportion of recognized outbreaks (i.e., the outbreaks in which at least one alarm is raised during their activity). Following Kleinman and Abrams [15] and Jafarpour et al. [12], we therefore use a ROC curve-like representation with the detection rate on the y-axis instead of the true positive rate, and use $dAUC_e$ to refer to the partial area under this curve. Figure 2 shows an example of the ROC-curve like representation and visualizes the area of $dAUC_{1\%}$. Kleinman and Abrams [15] proposed to use weighted ROC curves to also incorporate the influence of the measure timeliness (mean time to detect an outbreak). However, we argue that the weighing with the timeliness introduces a trade-off (importance of timeliness over detection rate) and a loss in interpretability of the absolute numbers.

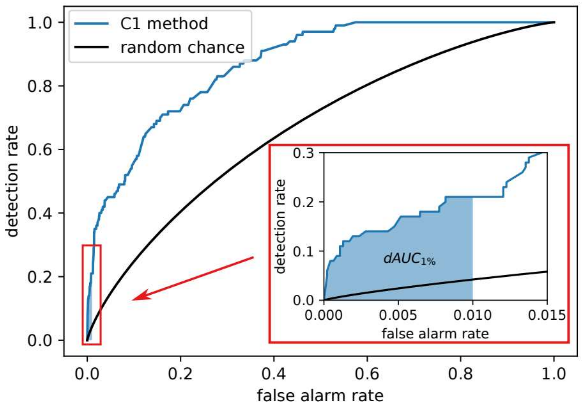


Figure 2: ROC curve using the detection rate on the y-axis. The better-than-chance performance is lifted above the diagonal since the detection rate is an interval-based metric.

6 EVALUATION

The key aspect of our experimental evaluation is to demonstrate that the fusion of p-values leads to a further improvement in performance compared to only using the binary output of the statistical algorithms. However, for a deeper understanding of our proposed approaches, we first performed experiments to evaluate the influence of including additional features for the stacking in Section 6.2, followed by an analysis of adapting the labeling for the learning in Section 6.3. Finally, using the obtained knowledge about the effect of the proposed techniques, we compare them with the underlying statistical algorithms in Section 6.4, which represents our main result.

6.1 Experimental Setup

As an implementation baseline for the statistical methods, we have used the R package *surveillance* [22] and adapted the implementation of the methods EARS (C1, C2 and C3), Bayes and RKI in order to also return p-values. All methods use the previous 7 time points as reference values, which is the standard configuration. For the ML part, we rely on the Python library *scikit-learn* [21]. To keep the evaluation simple, we use a *random forest* classifier. Basically, it learns an ensemble of randomized decision trees, which has proven to be robust in performance theoretically and practically [6, 27]. Each model is composed of 100 decision trees with a minimum number of instances per leaf of 5 and default settings otherwise. To allow comparability between the fusion methods, we also evaluated the approach which only combines the binary outputs of the statistical methods as proposed in [12, 24] and which we refer to as the *standard fusion*. Our preliminary experiments have shown that $\alpha = 0.5\%$ for the underlying statistical methods performs best for this fusion approach. For all evaluations, we focused on our proposed evaluation measure $dAUC_{1\%}$ where we fixed the constraint on the false alarm rate to be less than 1%.

Our evaluation is based on synthetic data which have been proposed by Noufaily et al. [19]. In total 42 different *test cases* are used which reflect a wide range of application scenarios allowing to analyze the effects of trend (T), seasonality (S1) and biannual seasonality (S2) explicitly. For each parameter configuration 100 time series are generated, each containing a total of 624 weeks. Following Noufaily et al. [19], the last 49 weeks of each time series serve as *evaluation data* which include exactly one outbreak whereas the first 575 weeks contain four outbreaks and represents the so called *baseline data*. Each outbreak starts at a randomly drawn week and the number of cases per outbreak is generated with a Poisson distribution with the mean equal to a constant k times the standard deviation of the counts observed at the starting week. The outbreak cases are then distributed over time using a log-normal distribution with mean 0 and standard deviation 0.5. We evaluated each stacking configuration separately for each test case using the baseline data of the 100 time series for training (in total 57.500 weeks including 400 outbreaks) and the remaining 4.900 weeks for testing (100 outbreaks), respectively. The statistical methods were applied separately for each time series in order to obtain the p-values as inputs for the learner as well as the predictions on the evaluation set.

Instead of reporting the average over $dAUC_{1\%}$ scores, which could have different scales for different test cases, we determined a ranking over the compared methods for each test case. Afterwards, each method's rank is averaged across the evaluated test cases to obtain an overall rank. In order to evaluate the effects of trend and seasonality explicitly, we average the rankings only over the test cases which include these effects. To differentiate between our proposed approaches, we use the notation $M(a, o, w)$ where $M \in \{P, S\}$ specifies whether p-value fusion (P) or the standard fusion (S) has been used, $a \in \{\neg\mu, \mu\}$ whether the mean is included, $o \in \{O_0, O_1, O_2, O_3\}$ which labeling for the learning, and $w \in \{0, 1, \ldots, 12\}$ the window size which has been used for the evaluation. In total, we tested 192 configurations from which we compare only a small subset , respectively, depending on the analyzed aspect.

6.2 Evaluation of Additional Features

The first aspect to review concerns the inclusion of the mean count over the last seven time points. Therefore, we have analyzed the effect of this feature independent of the other parameters using O_0 for the labeling of the outbreak and window size $w = 0$. The results for the average rank are displayed in Table 1. Comparing the standard to the p-value fusion method reveals a beneficial effect especially for the p-value approach, for which the variant including the mean achieves an average rank of 1.31 over 1.91. In contrast, the average ranks of 3.36 over 3.43 for the standard method not only shows that there are issues regarding the usage of the mean for some of the test case configurations, but also the substantial gap between using the binary outputs and the more fine-grained p-values. A closer examination reveals that the best improvement for both fusion methods can be achieved on time series without trend and seasonality. By adding effects like trend and seasonality, the mean changes over time, making it difficult for the learning algorithm to use this information. In contrast to the standard fusion, the p-value fusion method still enhances by including the mean over the previous time points.

The observation that the p-value fusion method is superior to the standard fusion can also be seen when comparing different window sizes. The results of this experiment, using O_0 for the labeling of the outbreak and not including the mean, are displayed in Table 2. In particular, no window configuration of the standard fusion method can outperform any of the p-value configurations with respect to the average rank. Overall, a window size of 1 performed best for both fusion approaches. Being able to compare to the most immediate previous output of the underlying statistical algorithms seems to make it easier to detect anomalies. In contrast, larger window sizes harm the overall performance, which suggests that the additional information is not relevant for detecting sudden changes and rather confuses the learner. Interestingly, on certain combinations of trend and seasonality a larger window size for the p-value fusion method seems to be beneficial. Actually, the increase of the window size also results in taking a further look back in the past allowing to detect effects like trend and seasonality achieving good results on the test cases which only contain biannual seasonality. However, the observed results for larger window sizes are inconsistent across the different test cases, making it difficult to draw valid conclusions.

6.3 Evaluation of the Labeling Adaptions

In addition to augmenting the input data, we have evaluated the effect of adapting the labeling of the epidemic period for the training of the stacking algorithm. The comparison shown in Table 3 was performed without the augmentation.

Table 1: Comparison of including or not including the mean in the data for ML algorithms: *overall* denotes all 42 test cases, $\{(\neg)T, (\neg)S1, (\neg)S2\}$ only cases (not) containing trend, annual/biannual seasonality, respectively. Each particular subset, fulfilling constraints on seasonality and trend, include 6 test cases.

Approach	Overall	$\{\neg T, \neg S1, \neg S2\}$	$\{\neg T, S1, \neg S2\}$	$\{\neg T, S1, S2\}$	$\{T, \neg S1, \neg S2\}$	$\{T, S1, \neg S2\}$	$\{T, S1, S2\}$
S($\neg\mu$, O_0, 0)	3.429	3.714	3.571	**3.000**	**3.429**	**3.286**	3.571
S(μ, O_0, 0)	**3.357**	**2.571**	**3.286**	3.571	3.571	3.714	**3.429**
P($\neg\mu$, O_0, 0)	1.905	2.571	1.857	2.000	1.714	1.714	1.571
P(μ, O_0, 0)	**1.310**	**1.143**	**1.286**	**1.429**	**1.286**	**1.286**	**1.429**

Table 2: Comparison of different window sizes for the data (including the mean and using the labeling O_0).

Approach	Overall	$\{\neg T, \neg S1, \neg S2\}$	$\{\neg T, S1, \neg S2\}$	$\{\neg T, S1, S2\}$	$\{T, \neg S1, \neg S2\}$	$\{T, S1, \neg S2\}$	$\{T, S1, S2\}$
S($\neg\mu$,O_0,0)	9.738	9.000	9.571	**8.286**	11.143	10.571	**9.857**
S($\neg\mu$,O_0,1)	**8.738**	**8.857**	**7.000**	9.000	**7.571**	**8.857**	11.143
S($\neg\mu$,O_0,2)	10.762	10.571	10.857	10.714	10.571	10.143	11.714
S($\neg\mu$,O_0,4)	11.310	11.429	11.714	11.714	10.857	12.000	10.143
S($\neg\mu$,O_0,6)	11.619	12.714	12.286	10.286	11.571	11.857	11.000
S($\neg\mu$,O_0,8)	11.548	11.143	11.571	12.000	10.857	12.000	11.714
S($\neg\mu$,O_0,12)	11.929	12.143	12.143	13.000	11.714	11.571	11.000
P($\neg\mu$,O_0,0)	5.000	5.714	5.000	5.714	4.429	3.714	5.429
P($\neg\mu$,O_0,1)	**3.405**	**3.143**	**2.571**	4.571	**3.286**	4.286	**2.571**
P($\neg\mu$,O_0,2)	4.381	5.000	4.714	4.000	4.571	4.571	3.429
P($\neg\mu$,O_0,4)	4.667	4.143	5.000	4.000	5.143	5.286	4.429
P($\neg\mu$,O_0,6)	4.310	5.000	4.429	3.857	3.857	3.857	4.857
P($\neg\mu$,O_0,8)	4.000	3.000	4.000	4.714	5.000	3.857	3.429
P($\neg\mu$,O_0,12)	3.595	**3.143**	4.143	**3.143**	4.429	**2.429**	4.286

Table 3: Comparison of the different labeling strategies for the epidemics (not using the average and $w = 0$).

Approach	Overall	$\{\neg T, \neg S1, \neg S2\}$	$\{\neg T, S1, \neg S2\}$	$\{\neg T, S1, S2\}$	$\{T, \neg S1, \neg S2\}$	$\{T, S1, \neg S2\}$	$\{T, S1, S2\}$
S($\neg\mu$, O_0, 0)	6.476	6.286	5.571	**4.857**	7.143	7.571	7.429
S($\neg\mu$, O_1, 0)	6.738	7.286	6.714	6.286	6.429	7.000	6.714
S($\neg\mu$, O_2, 0)	5.738	6.286	5.714	5.286	**5.429**	6.000	**5.714**
S($\neg\mu$, O_3, 0)	**5.524**	**5.286**	**5.143**	5.429	6.000	**5.429**	5.857
P($\neg\mu$, O_0, 0)	3.762	3.857	3.857	**2.714**	4.857	4.000	3.286
P($\neg\mu$, O_1, 0)	3.262	2.857	4.143	4.857	2.429	2.429	2.857
P($\neg\mu$, O_2, 0)	2.690	3.143	3.000	3.286	2.714	2.143	**1.857**
P($\neg\mu$, O_3, 0)	**1.810**	**1.000**	**1.857**	3.286	**1.000**	**1.429**	2.286

Table 4: Comparison of the standard fusion, the p-value fusion and each individual statistical algorithm.

Approach	Overall	$\{\neg T, \neg S1, \neg S2\}$	$\{\neg T, S1, \neg S2\}$	$\{\neg T, S1, S2\}$	$\{T, \neg S1, \neg S2\}$	$\{T, S1, \neg S2\}$	$\{T, S1, S2\}$
C1	5.381	6.429	5.429	4.143	5.714	5.714	4.857
C2	4.810	4.571	4.000	4.286	5.857	5.286	4.857
C3	4.690	5.429	4.571	4.286	4.857	4.429	4.571
Bayes	2.595	4.000	3.143	**2.571**	**1.571**	**1.714**	2.571
RKI	3.619	3.571	2.857	4.571	3.714	3.286	3.714
S(μ, O_3, 1)	5.238	3.000	6.000	5.714	4.714	5.857	6.143
P(μ, O_3, 1)	**1.667**	**1.000**	**2.000**	**2.429**	**1.571**	**1.714**	**1.286**

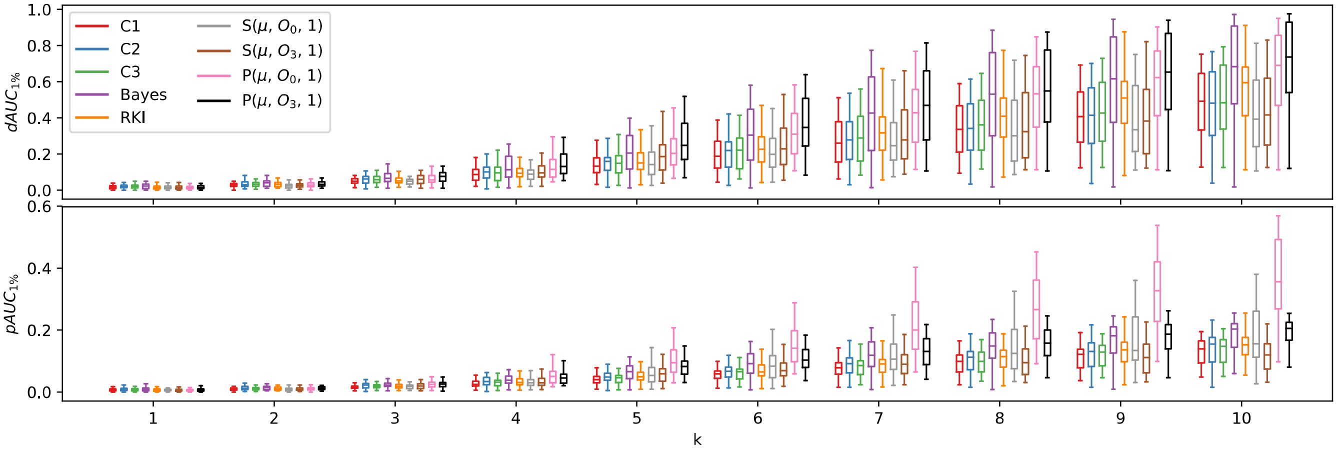


Figure 3: Results for the measures $dAUC_1\%$ and $pAUC_1\%$. Each box plot represents the distribution of measure values for a particular method computed over all 42 test cases for a fixed outbreak size defined by the parameter k (a bigger value for k indicate more cases per outbreak).

In general, we can observe that by narrowing the labeling of the outbreak on particular events (i.e., O_1, O_2 or O_3) a better performance can be achieved. This effect is clearly visible for the p-value fusion method and less obvious for the standard fusion method, for which the adaption O_1 seems to be an exception. In particular, learning only the peaks (O_3) achieved the best results for both fusion approaches. The benefit of this variant is that the learner can actually focus on the identification of strong and sudden peaks which is indeed the main goal of outbreak detection. However, in case of biannual seasonality the frequent change of the counts over the season results in many random peaks which apparently makes it difficult for the stacking approach to distinguish between an epidemic peak and a peak caused by random effects. On the test cases without trend ($\{\neg T, S1, S2\}$) outbreaks are better identifiable by also including the fading of the outbreak (O_0), whereas on the test cases which contain trend ($\{T, S1, S2\}$) the best option seems to be O_2, which only includes only the increasing counts until the peak of the outbreak is reached (O_2).

6.4 Comparison to the Statistical Surveillance Baselines

Considering the results of the previous experiment, we have chosen to evaluate the p-value and the standard fusion approach with a window size of 1, the adaption of the labeling O_3 and including the mean. In order to draw conclusions, we have evaluated the underlying statistical methods itself which serve as a baseline.

The results for the average rank are represented in Table 4. Here, we can observe that p-value fusion achieves the best rating across all test cases. In contrast, the performance of the standard fusion approach is often worse than the underlying statistical algorithms. In line with Texier et al. [24] and Jafarpour et al. [12], we can observe an improvement of the standard fusion approach on the time series without trend and seasonality. However, this improvement is not consistent for all compared test cases, resulting only in an average rank of 3.0 while our proposed p-value approach always achieves the best result. Indeed, the ability to detect outbreaks with the standard fusion approach is reduced since it is based on the output of the statistical algorithms given a particular pre-defined significance level α for them. This limits the information about sudden changes encapsulated in the training data which makes it pretty difficult for the ML algorithm to identify valuable patterns. A closer examination reveals that trend and seasonality has an impact on the evaluated stacking approaches. In particular, by learning over the baseline data of time series which include trend, the learner is fed with observations which are not representative for the future (evaluation data) due to the changed circumstances. Moreover, the learning algorithm usually assumes that the instances are considered to be independent and identically distributed in the learning data set, not allowing to capture concept drift. Our proposed approaches are not designed to adjust to these settings but we believe that further investigations on the influence of trend and seasonality and how they can be handled is an interesting avenue for future work.

Furthermore, we have evaluated the approaches with respect to the number of cases per outbreak. In contrast to the previous experiments, where the value for the parameter k (used to define the number of cases per outbreak) was randomly drawn between 1 and 10, we have fixed this parameter to a particular value for all time series of the 42 test cases. The results for the measure $dAUC_{1\%}$ across the 42 test cases with a fixed value for the parameter k is visualized as box plots, representing minimum, first quantile, mean, third quantile and maximum, in Figure 3. In addition to $dAUC_{1\%}$, we include the analysis of the $pAUC_{1\%}$ measure and compare to the original labeling O_0 in order to further investigate the effect of the labeling on detection rate and true positive rate.

As the cases per outbreak increases all methods are more likely to obtain a better performance. While the C1, C2, C3 and RKI method achieve comparable results across all outbreak sizes, we are surprised to observe that the Bayes method has a better performance in case of larger outbreaks. This contradicts our expectation that the RKI method should obtain the best results across these methods since the Poisson assumption was specifically used to generate the synthetic data. Regarding the p-value fusion approaches, the results confirm the better overall performance across all outbreak sizes while the performance of the standard fusion approach gets worse compared to the other methods with an increasing number of cases per outbreak. This gives further evidence that the standard fusion is not ideal. A closer examination of the graphs for the measures $dAUC_{1\%}$ and $pAUC_{1\%}$ reveals the difference between the adaption of the labeling for the learning. In particular, without adaption the ML algorithm achieves a tremendously better performance for the trade-off between the true positive rate and the false alarm rate. However, this also has an effect on the ability to detect outbreaks as discussed in Section 4.3, yielding a slightly worse result for the measure $dAUC_{1\%}$ than with adapting the labeling.

7 CONCLUSIONS

In this work, we introduced an approach for the fusion of outbreak detection methods using machine learning, more specifically stacking. The original idea is to use the *alarm* or *no alarm* prediction of the underlying statistical algorithms as inputs to the learner. We improved that setup by incorporating the p-values instead, which contain more information about the certainty of an event than the simple binary outputs. In addition, we proposed to add additional information to the learning data and to adapt the labeling of an outbreak in order to improve the ability to detect outbreaks. For evaluation, we proposed a measure based on ROC curves which better adapts to the specific need for a very low false alarm rate but still considers the trade-off with the detection rate.

Our experimental results on synthetic data show that the fusion of p-values improves the performance compared to the underlying statistical algorithms. Contrary to previous work, we could also observe that simple fusion of binary outputs using stacking does not always lead to an improvement. By incorporating additional information to the learning data, more specifically the mean count of the previous observations and the previous outputs of the statistical methods, the machine learning algorithm is able to capture more reliable patterns to detect outbreaks. Furthermore, the labeling of an outbreak has an influence on the performance for the classification algorithm to detect outbreaks. By setting the focus on the peak of an outbreak during the learning process, a better performance to detect sudden changes can be achieved.

The effectiveness of the proposed method has still to be confirmed on real data. Nevertheless, our results suggest that p-value stacking is generally well-suited for combining the outcomes of established methods for outbreak detection with only a low risk of decreasing performance. Moreover, stacking allows to enrich the detection by additional signals and sources of information in a highly flexible way. However, a major challenge remains the treatment of the outbreak annotations during training, since these labels are inherently non-binary (endemic vs. epidemic) and additionally noisy and unreliable for real data.

ACKNOWLEDGMENTS

This work was supported by the Innovation Committee of the Federal Joint Committee (G-BA) [ESEG projekt, grant number 01VSF17034]

REFERENCES

[1] G. Bédubourg and Y. Le Strat. 2017. Evaluation and comparison of statistical methods for early temporal detection of outbreaks: A simulation-based study. *PLOS ONE* 12(7):1–18.

[2] H. Burkom, L. Ramac-Thomas, S. Babin, R. Holtry, Z. Mnatsakanyan, and C. Yund. 2011. An integrated approach for fusion of environmental and human health data for disease surveillance. *Statistics in Medicine* 30(5):470–479.

[3] P. Chakraborty, P. Khadivi, B. Lewis, A. Mahendiran, J. Chen, P. Butler, E. Nsoesie, S. Mekaru, J. Brownstein, M. Marathe, and N. Ramakrishnan. 2014. Forecasting a moving target: Ensemble models for ILI case count predictions. In *Proceedings of the SIAM International Conference on Data Mining*. 262–270.

[4] D. Farrow, L. Brooks, S. Hyun, R. J. Tibshirani, D. Burke, and R. Rosenfeld. 2017. A human judgment approach to epidemiological forecasting. *PLOS Computational Biology* 13(3):1–19.

[5] T. Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27(8):861–874.

[6] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research* 15:3133–3181.

[7] R. Fricker Jr., B. Hegler, and D. Dunfee. 2008. Comparing syndromic surveillance detection methods: EARS' versus a CUSUM-based methodology. *Statistics in Medicine* 27(17):3407–3429.

[8] L. Hutwagner, T. Browne, G. Seeman, and A. Fleischauer. 2005. Comparing aberration detection methods with simulated data. *Journal of Emerging Infectious Diseases* 11(2):314–316.

[9] L. Hutwagner, W. Thompson, G. Seeman, and T. Treadwell. 2003. The bioterrorism preparedness and response early aberration reporting system (EARS). *Journal of Urban Health* 80(1):i89–i96.

[10] M. Jackson, A. Baer, I. Painter, and J. Duchin. 2007. A simulation study comparing aberration detection algorithms for syndromic surveillance. *BMC Medical Informatics and Decision Making* 7(1):6.

[11] N. Jafarpour, M. Izadi, D. Precup, and D. L. Buckeridge. 2015. Quantifying the determinants of outbreak detection performance through simulation and machine learning. *Journal of Biomedical Informatics* 53:180–187.

[12] N. Jafarpour, D. Precup, M. Izadi, and D. Buckeridge. 2013. Using hierarchical mixture of experts model for fusion of outbreak detection methods. *AMIA Annual Symposium Proceedings* 2013:663–669.

[13] M. Jordan and R. Jacobs. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6(2):181–214.

[14] B. Khaleghi, A. Khamis, F. Karray, and S. Razavi. 2013. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion* 14(1):28–44.

[15] K. Kleinman and A. Abrams. 2006. Assessing surveillance using sensitivity, specificity and timeliness. *Statistical Methods in Medical Research* 15(5):445–464.

[16] E. Lau, B. Cowling, L. Ho, and G. Leung. 2008. Optimizing use of multistream influenza sentinel surveillance data. *Journal of Emerging Infectious Diseases* 14:1154–1157.

[17] H. Ma, A. Bandos, H. Rockette, and D. Gur. 2013. On use of partial area under the ROC curve for evaluation of diagnostic performance. *Statistics in Medicine* 32(20):3449–3458.

[18] Z. Mnatsakanyan, H. Burkom, J. Coberly, and J. Lombardo. 2009. Bayesian information fusion networks for biosurveillance applications. *Journal of the American Medical Informatics Association* 16(6):855–863.

[19] A. Noufaily, D. Enki, P. Farrington, P. Garthwaite, N. Andrews, and A. Charlett. 2013. An improved algorithm for outbreak detection in multiple surveillance systems. *Statistics in Medicine* 32(7):1206–1222.

[20] A. Noufaily, R. Morbey, F. Colón-González, A. Elliot, G. Smith, I. Lake, and N. McCarthy. 2019. Comparison of statistical algorithms for daily syndromic surveillance aberration detection. *Bioinformatics*. In press.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

[22] M. Salmon, D. Schumacher, and M. Höhle. 2016. Monitoring count time series in R: Aberration detection in public health surveillance. *Journal of Statistical Software* 70(10):1–35.

[23] G. Shmueli and H. Burkom. 2010. Statistical challenges facing early outbreak detection in biosurveillance. *Technometrics* 52(1):39–51.

[24] G. Texier, R. Allodji, L. Diop, J. Meynard, L. Pellegrin, and H. Chaudet. 2019. Using decision fusion methods to improve outbreak detection in disease surveillance. *BMC Medical Informatics and Decision Making* 19(1):38.

[25] K. Ting and I. Witten. 1999. Issues in stacked generalization. *Journal of Artificial Intelligence Research* 10:271–289.

[26] D. Wolpert. 1992. Stacked generalization. *Neural Networks* 5(2):241–259.

[27] A. Wyner, M. Olson, J. Bleich, and D. Mease. 2017. Explaining the success of AdaBoost and Random Forests as interpolating classifiers. *Journal of Machine Learning Research* 18(48):1–33.

A Visual Analytics Framework for Analysis of Patient Trajectories

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

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

Ananth Kalyanaraman
Washington State University
ananth@wsu.edu

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

Rebekah Moehring
Duke University
rebekah.moehring@duke.edu

Bala Krishnamoorthy
Washington State University
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.

KEYWORDS

Electronic Health Records, TDA, Patient Trajectories

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 [5]. 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 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 data sets can be very useful in the study of hospital-acquired infections (HAIs) [12, 14], or for analyzing conditions such as sepsis [11].

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. [4]. 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 [16], 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 although the fraction of studies that focus on studying temporal trajectories have been relatively small. Giannoula et al. [3] presented a time-analysis framework 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. [6] presented a network representation of diseases to understand temporal 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. Nicolau et al. [15] used TDA to analyze breast cancer transcriptional data. They identified a unique subgroup of patients with 100% survival rate. Li et al. [9] 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.

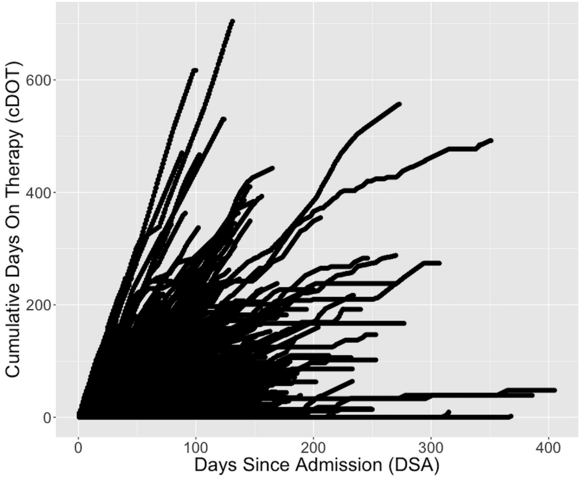


Figure 1: A scatterplot of patient trajectory data shown as a distribution of cumulative Days-of-Therapy (cDOT) values as a function of Day-Since-Admission (DSA). A single patient's trajectory of points is represented as a series of dots from DSA 1 to the last day of the patient's admission.

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:

Part 1: cDOT is responsive to a combination of temporal, spatial and treatment-based factors, although to varying degrees; and

Part 2: 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.

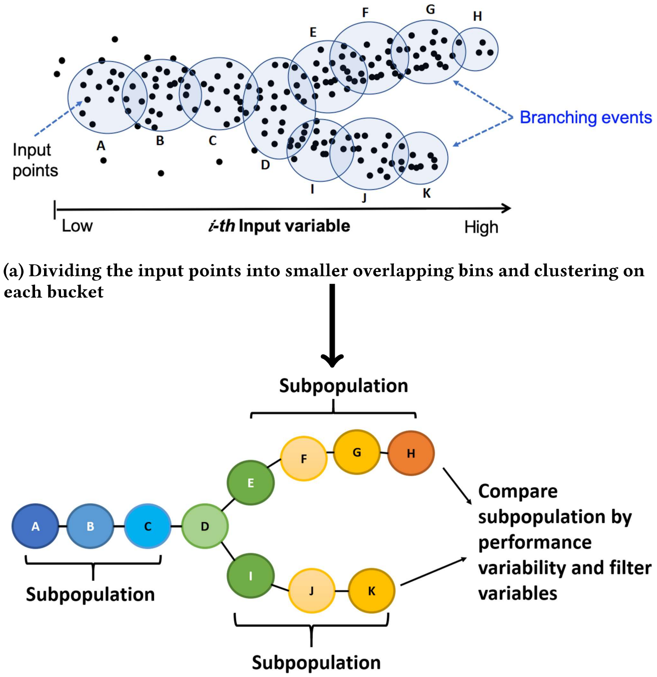
3.2 A Visual Analytics Framework

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, 13]. The important properties of topology that make it particularly effective for extracting structural features from large, high-dimensional data sets are: a) coordinate-free representation, b) insensitivity to small changes in data, and c) compressed representations [13]. 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 [10].

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 [7, 8], which is an implementation of the Mapper algorithm [16]. Hyppo-X is a computational tool for modeling and exploring multidimensional data where one set of (continuous) variables $f = \{f_1, f_2, \dots f_k\}$ 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.



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

(b) Compact topological object. Clusters are shown colored based on their originating interval over the input filter variable.

Figure 2: Generation of topological object from a point cloud.

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.

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 originating between two adjacent bins. We refer to the resulting graph as a *topological object* (simplicial complex, to be precise) as shown in Figure 2b. 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 [8].

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 2b 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 [7].

4 EXPERIMENTAL RESULTS

4.1 Data

We used the 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
Number of antibacterials used	66
Number of agent ranks	4

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 snapshot of the topological object output by our framework (in the actual tool, all topological objects allow interactive visualization). 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 patients 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. 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:

1) The majority of the clusters are dominated by patients in the adult medical/surgical ward (shown in red), followed by adult

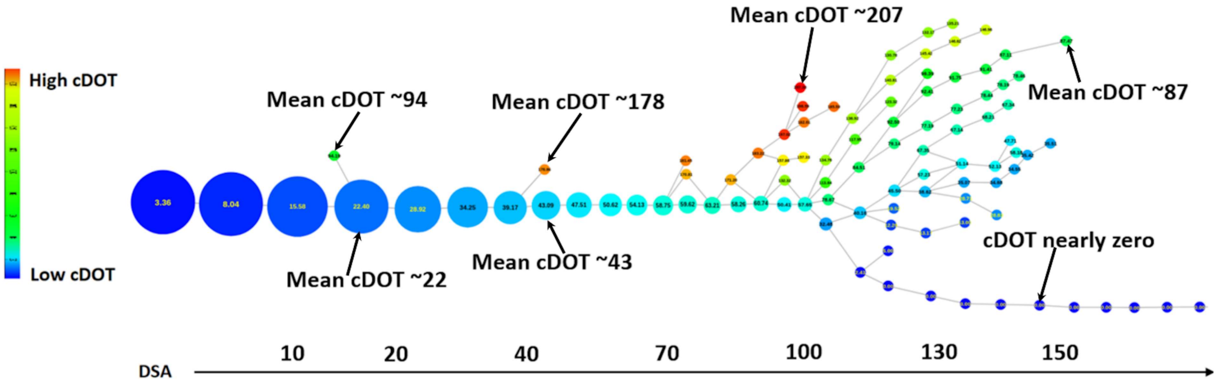


Figure 3: The topological object constructed using Days Since Admission (DSA) as a single filter function and cumulative Days On Therapy (cDOT) as the clustering attribute. The horizontal color bar indicates the gradient of DSA value from left (low) to right (hight). The node coloring and the labels 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.

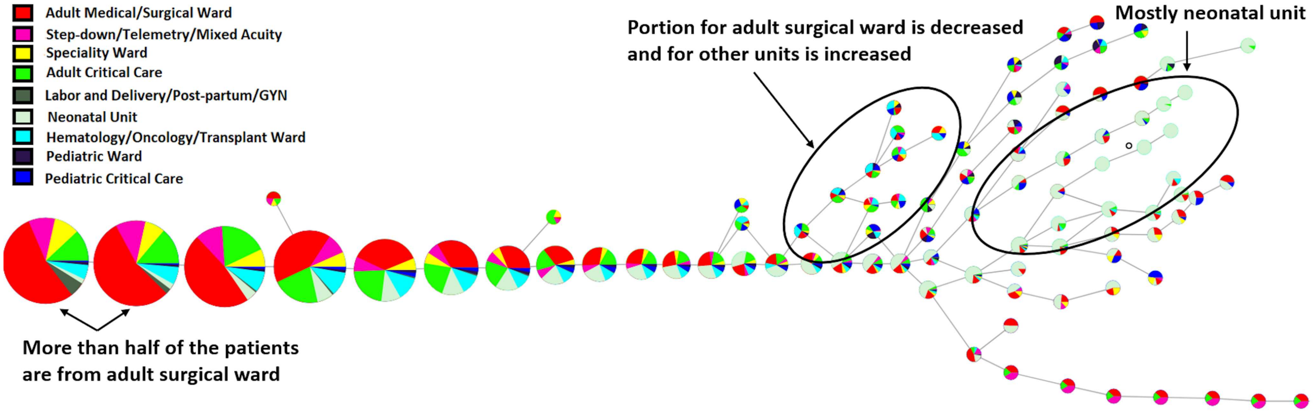

🟥 Adult Medical/Surgical Ward
🟪 Step-down/Telemetry/Mixed Acuity
🟨 Speciality Ward
🟩 Adult Critical Care
▨ Labor and Delivery/Post-partum/GYN
□ Neonatal Unit
🟦 Hematology/Oncology/Transplant Ward
■ Pediatric Ward
🟦 Pediatric Critical Care
Portion for adult surgical ward is decreased
and for other units is increased
Mostly neonatal unit
More than half of the patients
are from adult surgical ward

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

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.

2) 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).

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 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:

1) 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).

2) 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).

3) 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.

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. 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 6. 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 6:

1) 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.

2) Patients in the Duke Hospital are more likely (than ones in the other hospitals) to receive some antibiotic at least once during their stay.

3) The use of agent rank 4 is relatively more frequent at the Duke hospital than for the other hospitals.

4) 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. Also note that two flares can share an edge along a branch.

Figure 7 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:

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 7(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 7(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 7(B)). This subpopulation corresponds to patients mostly in either the Adult Medical/Surgical Ward or the Hematology/Oncology/Transplant Ward (see Figure 7(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

Topological data analysis has a potential to represent complex data in compact and visually friendly formats. In this paper, we have used this technique for clustering patient trajectories (by their antibiotic use or cDOT) so that they can be concisely viewed along the temporal dimension. We used multiple attributes such as hospital units or agent ranks to analyze and observe patterns in these clustered trajectories.

The analysis enabled us to find differing propensities for use of antibiotics within certain hospital units as well as 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

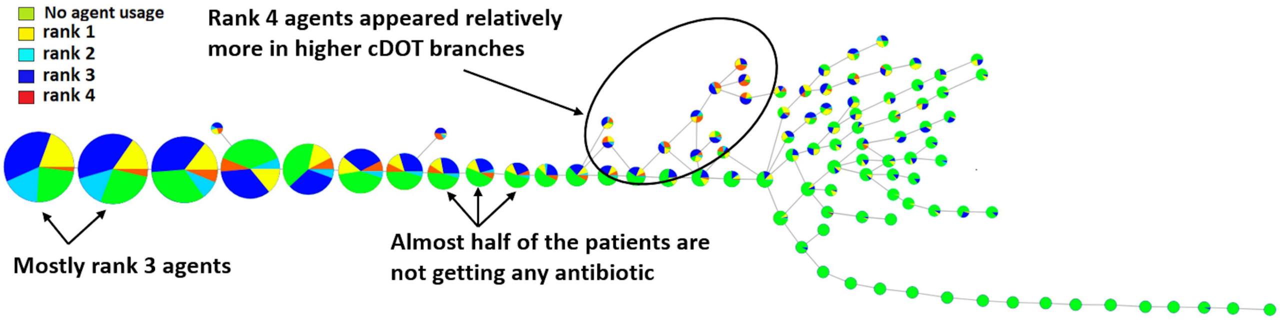
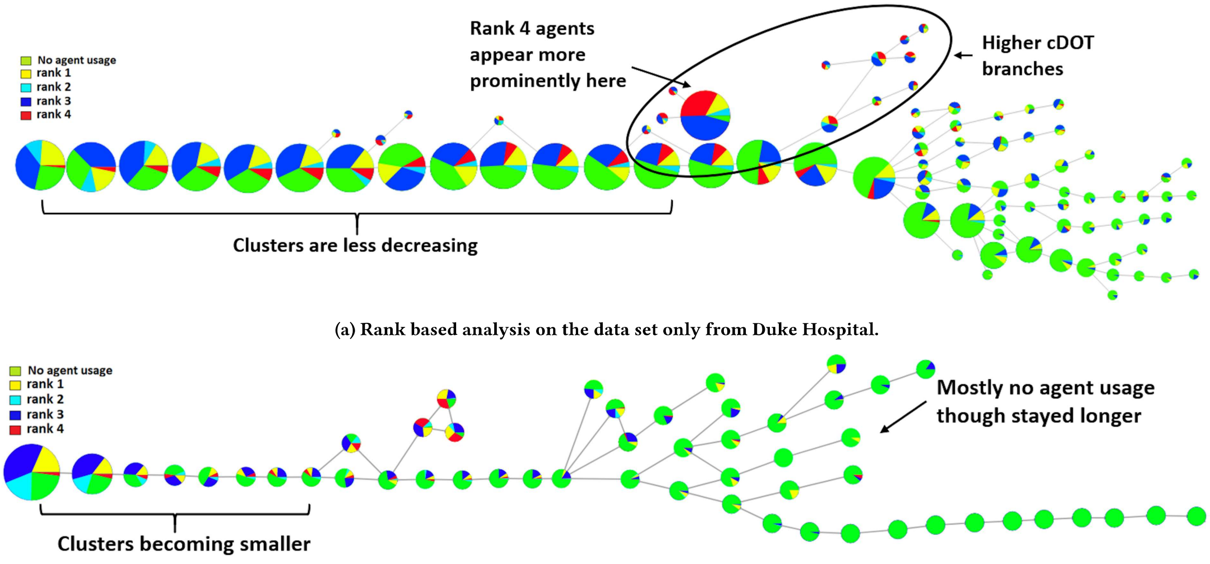


Figure 5: The pie chart in each node is representing the percentage usage of each agent group on that particular interval of time.



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

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

Figure 6: A comparison of rank based analysis on Duke (top) vs Non-Duke (bottom) Hospitals.

(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 work was in part supported by the U.S. Center for Disease Control and Prevention's (CDC's) investments to combat antibiotic resistance under award number 200-2018-96423; and by the CDC Cooperative Agreement RFA-CK-17-001-Modeling Infectious Diseases in Healthcare Program (MInD-Healthcare); and by the 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] 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.

[4] 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.

[5] 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.

[6] 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.

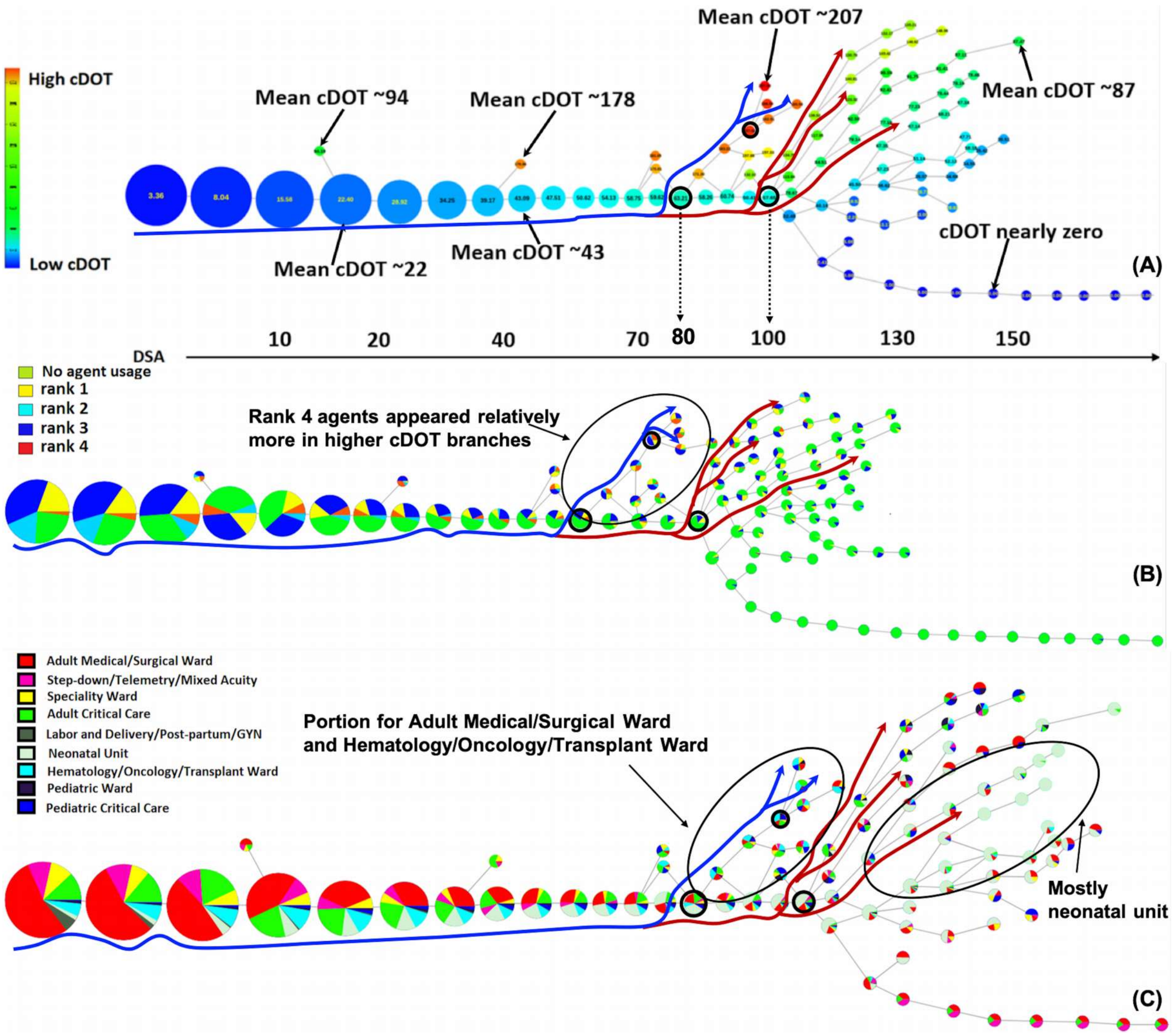


Figure 7: Topological object constructed using DSA as a single filter function (shown earlier in Figure 3), now also showing the interesting flares detected by our method. The nodes are arranged from left to right with chronological order of mean DSA values. (A) Each cluster colored by its mean cDOT, 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.

Nature communications 5 (2014), 4022.

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

[8] 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.

[9] 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.

[10] 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).

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

[12] 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.

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

[14] 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.

[15] 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.

[16] Singh, G., Mémoli, F., and Carlsson, G. E. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG* (2007), pp. 91–100.