2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)

Big Data Techniques for Public Health: A Case Study

Yannis Katsis[†], Natasha Balac[¶], Derek Chapman[‖], Madhur Kapoor[†], Jessica Block[¶], William G. Griswold[†], Jeannie Huang[‡], Nikos Koulouris[†], Massimiliano Menarini[†], Viswanath Nandigam[§], Mandy Ngo[†,*1], Kian Win Ong[†,*2], Yannis Papakonstantinou[†], Besa Smith[‡], Konstantinos Zarifis[†] Steven Woolf[‖], Kevin Patrick[‡,¶]

[†]Dept. of Computer Science and Engineering, [‡]Department of Family Medicine and Public Health, School of Medicine, [§]San Diego Supercomputer Center, [¶]The Qualcomm Institute
University of California, San Diego
La Jolla, CA USA
&
[‖]Center on Society and Health, Virginia Commonwealth University
Richmond, VA, USA

ikatsis@cs.ucsd.edu

***Abstract* — Public health researchers increasingly recognize that to advance their field they must grapple with the availability of increasingly large (i.e., thousands of variables) traditional population-level datasets (e.g., electronic medical records), while at the same time integrating additional large datasets (e.g., data on genomics, the microbiome, environmental exposures, socioeconomic factors, and health behaviors). Leveraging these multiple forms of data might well provide unique and unexpected discoveries about the determinants of health and wellbeing. However, we are in the very early stages of advancing the techniques required to understand and analyze big population-level data for public health research.**

To address this problem, this paper describes how we propose that big data can be efficiently used for public health discoveries. We show that data analytics techniques traditionally employed in public health studies are not up to the task of the data we now have in hand. Instead we present techniques adapted from big data visualization and analytics approaches used in other domains that can be used to answer important public health questions utilizing these existing and new datasets. Our findings are based on an exploratory big data case study carried out in San Diego County, California where we analyzed thousands of variables related to health to gain interesting insights on the determinants of several health outcomes, including life expectancy and anxiety disorders. These findings provide a promising early indication that public health research will benefit from the larger set of activities in contemporary big data research.

Keywords—public health, big data, machine learning, data exploration, data visualization.

I. INTRODUCTION

Public health research is typically done by focusing on a small set of indicators (i.e., variables) that are suspected to be associated with a particular health outcome. For instance, the authors of [7] hypothesized that there is a correlation between maternal residential proximity to major roads and low birth weight. In order to test this hypothesis, they combined and analyzed several variables related to the stated hypothesis; in this case distance from major roads, temperature, air pollution, and noise data.

However, the increasing availability of large population-level data related to health (such as all-payer claims databases, vital records, electronic medical records, and sensor data) and to varying aspects of health (e.g., environmental exposures, demographics, socioeconomic factors, and consumer purchasing behaviors), has led public health researchers to explore novel approaches to data aggregation and analysis. There is a growing recognition that leveraging such population-level datasets with thousands of variables (which for the purpose of this paper we refer to as "big data")[1], can lead to important and unexpected public health discoveries [14][23][25]. For instance, imagine leveraging a large dataset on environmental data to study birth weight. Using such a dataset one could produce evidence not only for association of the birth weight with major roads, but also for associations with other factors that one may have never suspected (such as proximity to parks, etc.).

However, little has been done to advance techniques for leveraging big population-level data for public health studies[2]. How do big data public health studies differ from

*1 Author currently at Space and Naval Warfare Systems Center, Pacific. The work was done while the author was working at UC San Diego.
*2 Author currently at Facebook, Inc. The work was done while the author was working at UC San Diego.

[1] There have been varying definitions of "big data", referring among others to large volumes of data, large data generation rates, or significant heterogeneity. For the purpose of this paper, big data refers to a large number of variables (in the order of thousands), since this significantly changes the nature of data analysis required compared to a smaller number of variables.

[2] Even though there has been a large body of work on big data analytics for health, most of it has focused on genomics, proteomics and related systems biology issues. Other efforts have focused on large-scale analysis of electronic medical record data, aggregations of images from pathology or functional magnetic resonance imaging data. Very little attention has been given to the unique requirements of public health studies, which is

978-1-5090-4722-2/17 $31.00 © 2017 IEEE
DOI 10.1109/CHASE.2017.81

traditional public health studies? Can we analyze big datasets using standard statistical techniques employed in traditional public health studies or do we need to adopt other data analytics approaches?

In this paper, we study the problem of leveraging big data for public health discovery. We show that there are two types of big data studies; studies that are driven by hypotheses (similarly to traditional epidemiologic studies), but also novel studies that are driven from the data instead (where the researcher does not have a particular question in mind but is instead exploring the dataset for potentially important patterns). We show that in both types of big data studies, traditional statistical techniques are not sufficient, as they either do not scale to big datasets (in the case of hypothesis-driven studies), or they do not support the particular type of study altogether (in the case of data-driven studies). To address this problem, we leverage two broad classes of big data analytics techniques to enable the two types of studies: We use machine learning techniques to enable hypothesis-driven studies and data visualization techniques to enable more open-ended data-driven studies. In each case we propose specific novel techniques to answer common public health questions.

Our techniques were informed and are presented in the context of an exploratory big data case study carried out in San Diego. Conducted as a collaboration between UC San Diego, the Center on Society and Health at the Virginia Commonwealth University, and San Diego's Health and Human Services Agency (HHSA), this study explored the determinants of several health outcomes by leveraging thousands of variables related to the health of the residents of San Diego County, including socio-economic, environmental, behavioral and traditional public health data. While the presented techniques were developed in the context of this study, we believe that they apply more generally to many areas of public health research.

The paper is structured as follows: We start in Section II by presenting the state of the the art in public health studies. In Section III we identify the challenges that arise when we attempt to use traditional techniques to process big datasets and in Section IV, we discuss how we addressed these challenges in our case study through the use of big data techniques. Finally, in Section V we discuss lessons learned from our case study and areas of future work.

II. State of the Art in Public Health Studies

Public health is defined by the Institute of Medicine as "what we, as a society, do collectively to assure the conditions in which people can be healthy" [11]. While medicine is concerned with the treatment of disease in individuals, public health emphasizes prevention of disease and promotion of health and well-being. The focus of public health is also on communities rather than individuals. Federal, state and local governments are responsible for carrying out three core functions of public health: assessment (collection and data analysis), policy development, and assurance (linking people to services) [11]. Public health officials rely heavily on assessment data about the populations or communities they serve in order to set priorities, develop policies, and plan, implement and evaluate intervention programs.

Epidemiology is the basic science of public health that provides the information needed to guide public health policy and practice. The ultimate goal of epidemiologic studies is to control the occurrence of diseases, disability, and other adverse health conditions through prevention (e.g., modify or remove the causes of disease) or by improving the health status of populations. There is, in fact, a long history in public health of effective preventive measures being introduced long before the discovery of the causative or preventive agent [30]. Successful examples of this include the dramatic decreases in infectious disease and subsequent enhanced longevity over the past three centuries due to improvements in nutrition, sanitation, and hygiene long before vaccines and antibiotics were discovered [17]. More recent examples are the identification of the link between smoking and lung cancer [8] and prone sleeping position and sudden infant death syndrome [16].

Most epidemiologic studies that are used to guide decisions in public health practice fall into two broad categories: observational studies and descriptive studies[3]. Observational studies rely on the observation of individuals, while descriptive epidemiologic studies rely on population-level data to provide an understanding of the characteristics of who is at risk, in which places disease rates are highest, and temporal patterns of disease. While the lack of individual-level factors in the descriptive studies limits their ability to make causal inferences, policymakers are often specifically interested in population-level influences on health. For example, to prevent lung cancer, one prevention approach could be for doctors to encourage individual patients to stop smoking. A more efficient approach, however, may be to identify and address factors at the population level that may influence the smoking behaviors of individuals (e.g., advertising, cigarette taxes, poverty-related stress). In this paper we focus on descriptive studies as they are more amenable to the use of population-level datasets that are easily accessible (in contrast to individual-level data that are often hard to obtain, due to privacy concerns).

Regardless of the particular topic of interest, the process of conducting descriptive research in public health practice follows a standard series of steps [2]: statement of the problem, review of relevant literature, formulation of the study question or hypothesis, sample selection, selection and measurement of indicators of exposure and outcome, and data analysis including evaluation of the role of chance and bias. A key component of the entire process is the hypothesis generation. Each study is focused on proving a particular hypothesis formulated by the epidemiologist.

To check the validity of a hypothesis, epidemiologists commonly use standard statistical techniques. Since the most common questions in public health are binary (e.g., presence of disease or not, risk factor present or not, qualify for service or not), the most commonly used measures of

the focus of this work. For a survey of big data approaches for healthcare, the reader is referred to [9][20].

[3] For a complete review of epidemiological studies, including other types of studies, the reader is referred to [10][21].

effect are relative measures (the risk ratio for common outcomes and the odds ratio for rare conditions) that compare the rate of a health outcome in groups "exposed" to a risk factor relative to an "unexposed" or comparison group. Thus, binary logistic regression modelling using a set of variables based on known risk and potentially confounding factors in the literature is the most commonly used statistical technique when epidemiologists are addressing research questions relevant to public health practice. When the outcome of interest is continuous (e.g., life expectancy), a similar process of literature-based variable selection is employed in the context of a multiple linear regression analysis.

III. Challenges In Big Data Studies

At first glance, one may think that epidemiologic studies that use a large number of indicators/variables (in the order of thousands) can be carried out in the same way as traditional epidemiologic studies that employ a limited number of variables. In reality, however, this is far from trivial for a variety of reasons:

Scale statistical techniques to large amounts of data. Statistical techniques commonly used in traditional epidemiology do not scale to large numbers of indicators. For instance, to compute how combinations of indicators affect a health outcome, epidemiologists commonly use multiple regression analysis. While this works for a limited number of indicators, it becomes increasingly hard to keep an overview of the result when the number of indicators increases. To solve such a problem in the presence of thousands of indicators one has to resort to non-traditional big data analytics techniques. However, even deciding which big data analytics technique to use is challenging as we explain next.

Identify big data analytics techniques that can answer common epidemiological questions. Because of the novelty of using big data techniques in public health, there is little precedent about which techniques to use. The machine learning literature contains a large variety of techniques, each suitable for different problems. Examples include decision trees, support vector machines (SVMs), neural networks, association rule mining, clustering and others. Moreover for each type of technique, there are multiple algorithms that can solve the problem. For instance, clustering methods are classified into partitioned, hierarchical, density-based clustering etc. To successfully use data analytics techniques in an epidemiologic study, it is imperative to first, understand which are the main classes of questions of interest to public health researchers, and second, identify the particular machine learning algorithms that fit each such class of questions.

From hypothesis-driven to data-driven studies. Finally, one has to also account for the different nature of studies that use big data. As explained above, traditional epidemiologic studies start from a specific hypothesis (e.g., that birth weight is influenced by the proximity of the maternal residence to major roads [7]). Then the entire study, including the data collection and the analysis are structured around this hypothesis. While it is possible to use such a hypothesis-driven approach when working with big amounts of data, such data can also be used for a new type of open-ended data-driven exploratory studies, in which one integrates a large number of indicators covering different aspects of human health (e.g., environmental, socioeconomic, behavioral, etc.) and tries to discover interesting patterns without having a specific question in mind. Since traditional studies have been hypothesis-driven, explorations of new approaches are in order to allow epidemiologists to carry out data-driven studies.

IV. A Big Data Case Study

To explore how big amounts of population-level data can be leveraged to make interesting public health discoveries, we worked on a case study centered on public health issues in San Diego County, California. The choice of location was made primarily for two reasons: First, the ease of getting access to large datasets, since it is the county where UC San Diego is located. Second, the diversity of the county, which makes it especially interesting for public health researchers: San Diego County's location (being close to the US border with Mexico and covering a large area from the Pacific Ocean coast to the desert), magnitude (being the fifth most populous county in the US), and population characteristics give it a unique environmental, ethnic, and socioeconomic diversity.

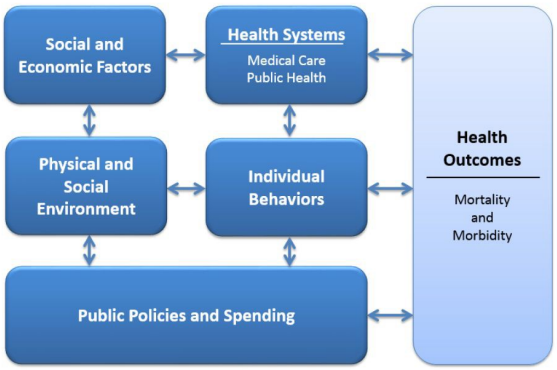

Figure 1. High-level grouping of determinants of health

To bootstrap our study, we identified and integrated a large number of representative data (in the order of thousands of indicators) covering the high-level groups of factors that are known to affect our health (shown on Figure 1, which was adapted from [29]): social and economic factors (such as education and income), physical and social environment (such as traffic density and air pollution), individual behaviors (such as smoking, exercising, and consumer buying patterns), health systems (such as insurance status), and health outcomes (such as hospitalization and emergency department visits for different conditions)[4].

[4] Our study did not include data on public policies/spending, as policies are usually broad, resulting in the lack of data at the sub-county level.

Since different datasets were provided at different geographic granularities, we ended up with two sets of integrated data: The first dataset contained 3,818 indicators at the level of the subregional areas (SRAs) (of which there are 41 in the San Diego County). While this dataset contained important health outcome information (i.e., hospitalization and emergency department visit data for different conditions), its geographic granularity was restricted due to privacy reasons. Therefore, we also created a second dataset that contained 22,712 indicators at the level of census tracts (of which there are 628 in the San Diego County). Figure 2 shows the data that were integrated into each of the two datasets.

Data Source	Indicator Count
Subregional area (SRA)-level dataset	**3,818**
HHSA Behavioral Health Data (Hospitalizations & Emergency Department visits for behavioral health conditions)	1,170
HHSA Demographics (Demographics)	300
ESRI Market Potential Data (Consumer buying patterns and behaviors)	2,234
SANDAG Healthy Communities Atlas (Data on physical and built environment)	114
Census-tract level dataset	**22,712**
American Community Survey 2012 (5-Year Estimates) (Census demographics)	22,547
CalEnviroScreen 2.0 (Pollution data)	45
Life Expectancy Data	6
SANDAG Healthy Communities Atlas (Data on physical and built environment)	114

Figure 2. Contents of the two integrated datasets used in the case study

To analyze the data we experimented with two broad classes of big data analytics techniques that cover the two ends of the spectrum between targeted hypothesis-driven discovery and open-ended data-driven exploration: To answer specific questions, such as computing the factors that affect the life expectancy of the county's residents, we used traditional data analytics techniques, borrowed from the machine learning literature. To allow more open-ended discoveries we implemented a visual data exploration platform, that allows public health researchers to visually explore the data and their correlations. We next describe each of these techniques in detail, starting with machine learning-based data analytics.

A. Machine Learning-based Data Analytics

Machine learning is a field of computer science that provides computers with the ability to learn without being explicitly programmed. The notion of machine learning is to automate the creation of analytical models by enabling algorithms to learn continuously from complex data. Machine learning techniques have been found very effective and relevant to many real world applications in bioinformatics, healthcare, marketing, sales, banking, finance, transportations, etc. As we show next, machine learning techniques have the potential to also greatly help public health researchers, by allowing them to leverage existing datasets to gain interesting insights into the determinants of health outcomes.

In this Section we present the major public health tasks that can be enhanced through the use of machine learning techniques, as identified during our case study. Although we report the result of using these techniques on the particular case study, the main focus of this Section is on how machine learning can enhance common public health tasks that arise in a variety of public health studies.

Using feature selection to compute the top determinants of health outcomes. Given a large number of indicators, one of the most interesting questions from a public health perspective is figuring out which indicators are the most important determinants of a particular health outcome. For instance, in our case study public health researchers were interested in finding out of all the indicators which contribute most to life expectancy. While in traditional public health studies this is done by creating linear regression models and studying the percentage of the variance accounted for in the model, this approach does not easily scale to large number of variables. To solve this problem, we use feature selection techniques, as described next.

The main objective of feature selection is to select a subset of most important features from multidimensional datasets. The machine learning literature contains a large variety of feature selection techniques, such as statistical measure-based filter methods, wrapper methods, and regularization-based embedded methods. In our use case we got promising results through a combination of the Random Forest approach and the Extra-trees classifier[5]. A Random Forest is an ensemble-based method that works by creating numerous decision trees [5][19] in a random fashion so as to avoid overfitting. Similarly, an Extra-trees classifier is a variant of the Random Forest classifier with more randomness built-in. To produce the desired model, we combine both techniques. In particular, given a specific health outcome, we build two models predicting the outcome; one using the Random Forest classifier and another the Extra-trees classifier. Each model contains the most important indicators that influence the chosen health outcome. The final list of indicators that affect the health outcome the most is produced by taking the intersection of the indicators chosen by both models. This is accomplished through a bagging-based approach [6].

In our study we utilized this technique to compute the top determinants of life expectancy. While it is widely recognized that socioeconomic factors, such as income and social status, have a strong influence on life expectancy, the degree to which the environment (both the physical and the built environment) affects life expectancy is less well-understood. To examine this relationship between environment and life expectancy, we used feature selection to find the top environmental factors that affect life expectancy based on the census tract-level dataset. Figure 3

[5] A comparison of the applicability of different feature selection algorithms for public health studies will be the focus of our future work.

shows the list of the top 8 such indicators. Interestingly, it is shown that based on the integrated data, the average number of crimes, the grocery store density, the fast food outlets and the sidewalk length are significant determinants of life expectancy. Additionally, the Machine Learning predictive models were able to reveal several interesting insights including that low life expectancy is characterized by large urban areas and lack of sidewalks. Furthermore, the areas with high life expectancy have high community engagement and access to walkable areas.

Avg. Violent Crime rate
Total park acreage
Fast food density per square mile
Block groups with minority areas
Composite Score: Pedestrian Traffic Safety
Percentage of block groups that intersect with parks
Percentage of sidewalk coverage
Count of grocery stores locations

Figure 3. Top environmental factors related to life expectancy

Evaluating the effectiveness of feature selection. To evaluate the effectiveness of the feature selection algorithm, we compared the features selected by the algorithm to the most important features that a domain expert manually chose by hand. Since the effect of environmental factors (shown above) on life expectancy is not yet well understood by experts, for this comparison we excluded the environmental factors and had the expert select according to this domain knowledge all the remaining indicators that are important determinants of life expectancy from a theoretical standpoint. This was a long and tedious process that yielded 1,045 manually chosen indicators (out of 16,384 census tract-level indicators[6]).

To compare the features selected by the domain expert to those selected by the feature selection algorithm, we use the two feature sets to create two corresponding models for life expectancy prediction (the models were built using Random Forest and Extra-trees classifiers combined through bagging). We then compared the accuracy of the two models. The accuracy of a model, defined as the ratio of correctly identified points to the total number of points, was computed by taking the ratio of life expectancy classes (low, low-medium, medium, medium-high, and high) correctly classified to the number of census tracts. The manual approach yielded an accuracy of 75%, while the automatic approach yielded an even higher accuracy of 77%. However, the accuracy of a model may not tell the entire story, because a model with high accuracy may predict well classes that appear very frequently (such as medium life expectancy), but fail to correctly classify other classes. To check whether this is the case, we examined the confusion matrix of the automatically inferred model. The confusion matrix, shown in Figure 4, explains how the model predicts all life expectancy classes. As we can see, the model correctly predicts the under-represented and most important low and high life expectancy classes. Moreover, it does not produce a significant number of false negatives/positives.

Finally, we investigated whether we could improve the model by combining the results of both the manual and the automatic approach. To this end, we build a new model by taking the union of the indicators chosen by either of the approaches. The resulting hybrid indeed outperformed the individual models, resulting in an accuracy of 80%.

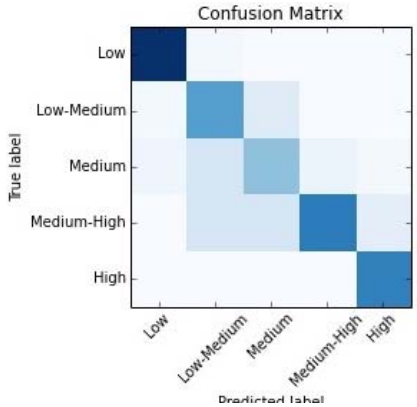

Figure 4. Confusion matrix depicting how well the feature selection algorithm predicts life expectancy

Summing up, automatic feature selection looks very promising, as it is not only faster than having a domain expert manually find all factors that affect a certain health outcome, but it yields accuracy that is higher than that of the domain expert when the indicators are well understood by the expert. Additionally it can also provide novel insights not predicted by the expert, when the effect of the indicators on the health outcome is not well understood (as is the case with the environmental indicators discussed above).

Using association rule mining to compute interesting associations. While the top determinants of a health outcome are important to understand which factors affect the health outcome, they do not explain how different factors interact with each other to affect the outcome. For instance, while smoking, drinking, sedentary behavior and non-healthy eating are all risk factors for decreased life expectancy, it is their combination that has the most detrimental effect [15]. Traditionally, this study of how combinations of factors affect a health outcome has been done using multiple regression analysis (i.e., an analysis where the value of the health outcome is predicted by a combination of multiple indicators). While such techniques work for the typically limited amount of variables that are involved in traditional public health studies, they do not scale up to thousands of variables, since it is extremely hard to interpret the results. To address this problem, we study the interaction between multiple indicators and a health

[6] Note that the study did not consider all 22,712 census tract level indicators as some non-interesting indicators (i.e., those with low variance) were removed in pre-processing.

outcome by leveraging existing algorithms for association rule mining.

Association rule mining is a rule-based machine learning method for discovering interesting relationships between variables in large datasets. It is intended to identify strong rules discovered in data using some measures of interestingness [18]. Association rule mining came into existence as market basket analysis on nominal datasets. Based on the concept of strong rules, association rules were designed for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets [1]. Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. The support of a rule is defined as the percentage of regions that satisfy the left-hand side of the rule in the dataset, while the confidence is defined as the measure of how often the right-hand side of the rule is found to be true within regions satisfying its left-hand side.

Association rule mining techniques include Apriori, FP-growth, Eclat, etc. For mining association rules in a public health setting we select the Apriori algorithm, primarily due to its simplicity. The algorithm works by identifying frequently occurring patterns in the dataset with high confidence and support.

Violent Crime Rate (Moderate) AND Median Household Income (Low) => Anxiety Disorder Rate (High)
Violent Crime Rate (Low) AND Percent population with no high school education (Low) => Anxiety Disorder Rate (Low)
Median Household Income (Low) AND Percent Hispanic population (High) => Anxiety Disorder Rate (High)
Avg. violent crime rate (Low) AND Percentage of population that car-pooled to work (Low) => Anxiety Disorder Rate (Low)
Violent Crime Rate (Medium) AND Median Household Income (Low) AND Percent Hispanic population (High) => Anxiety Disorder Rate (High)
Buys fat-free food (Medium) AND Consumes Nutrition focused diet (High) => Anxiety Disorder Rate (Medium)

Figure 5. Apriori rules for Anxiety Disorder Rates

In our case study we used association rule mining to discover interesting patterns of indicators that affect emergency department discharge rates for anxiety disorders in the San Diego County. Figure 5 shows the top 6 association rules inferred from the SRA-level dataset. Each rule shows an interesting association inferred from the data. For instance, the first rule shows that areas with moderate violent crime rates and low median household income have high emergency department discharge rates for anxiety disorders.

Leveraging association rule mining to move towards precision public health. Interestingly, in addition to making it possible to scale the analysis of the effect of combinations of variables to thousands of indicators, association rule mining also has an additional important benefit for public health studies. In contrast to traditional regression analyses that try to find a model that predicts the health outcome for the entire population, association rule mining produces rules that may hold only for a subset of the regions. This behavior of association rule mining algorithms can have very important repercussions in public health, as it allows public researchers to discover and study patterns that hold in a subset of the population, allowing them to create more targeted interventions for the particular sub-population. Conceptually, this can be thought of as the equivalent of precision medicine (which promises extremely targeted healthcare for individuals) for public health.

Using clustering to identify sub-populations with similar characteristics. Interestingly, it is not only association rule mining that allows public health researchers to move their focus from the entire population to sub-populations. By examining the needs of public health researchers and machine learning techniques, we discovered that other machine learning techniques can also be used to answer questions regarding sub-populations. In particular, one such question of considerable importance to public health researchers is identifying sub-populations that behave in a similar way w.r.t. a given set of indicators. This is in general important for two reasons: First, it helps public health officials target limited resources and/or create individualized interventions for different sub-populations, as we explained above. Second, finding regions that behave similarly is also useful for identifying the determinants of a health condition. For instance, finding which are the two groups of regions with the worst and best behavioral health outcomes, respectively allows researchers to compare and contrast the two groups to identify what makes the regions with the best outcomes different from the ones with the worst outcomes. To support these scenarios and allow researchers to find groups of regions that behave similarly w.r.t. a set of indicators, we leverage clustering techniques, suggested in the machine learning literature.

Clustering is an unsupervised technique that attempts to group objects to optimize the criterion that states that distance among objects in the same cluster is minimized and distance among objects in different clusters is maximized [27]. Various proximity measures can be used for computing the distance between the pairs of objects, such as Euclidean, Cosine, and Manhattan distance. In traditional clustering, all the features are used while computing the distance between a pair of objects over the entire training dataset of features.

Clustering methods are mainly classified into partitioned clustering, hierarchical clustering, density-based clustering, graph theoretic clustering, soft computing-based clustering, and matrix operation-based clustering [12]. While each class of clustering algorithms has its own unique advantages, one of the main requirements in a public health setting is that the selected algorithm produces results that can be easily

communicated and interpreted by public health researchers. To achieve this ease of exposition of the results, our analysis adopted partitioned clustering methods; i.e., clustering algorithms that assign each object (in our case a region) to *exactly one* of the k clusters, where k is a parameter. One of the most popular clustering algorithms is *k*-means [13]. *k*-means assigns objects to the nearest cluster centroid (i.e., cluster center) iteratively until there is no more assignment possible. However, *k*-means uses in general as cluster centroids new artificially created data points that may not exist in the original dataset. For instance, when clustering regions based on a set of features, *k*-means uses as cluster centroids new virtual regions that have certain values for the features, even though these regions may not exist in the input. Based on our requirement of interpretability of the results, we adopted in our analysis the *k*-medoids [13] clustering algorithm, which is similar to *k*-means but creates cluster centroids using existing data points (in our case regions) instead.

Another common differentiator between clustering algorithms is whether the number *k* of clusters to be created is selected by the user or automatically identified by the algorithm. While in some cases, a public health researcher may want a specific number of clusters (e.g., if she wants to categorize regions as "good" and "bad" w.r.t. a certain health outcome, thus necessitating the use of k=2), we found out that in most cases researchers are not as much interested in the exact number of clusters, as in the knowledge that these clusters are coherent (i.e., that regions within a cluster have a high degree of similarity, while regions across clusters have a high degree of dissimilarity). To this end, we used the silhouette score [22] to select the number of clusters produced by the *k*-medoids algorithm. The silhouette score is a similarity metric ranging from -1 to 1 signifying the similarity of an object to other objects in its own cluster as compared to objects in other clusters. To incorporate the silhouette score into the clustering process, we ran the *k*-medoids algorithm for different values of *k* and selected the clustering with the highest silhouette score. As a result, public health researchers are able to simply choose the indicators that they want to use as the basis for the clustering and get as a result a set of coherent clusters w.r.t. the chosen set of indicators.

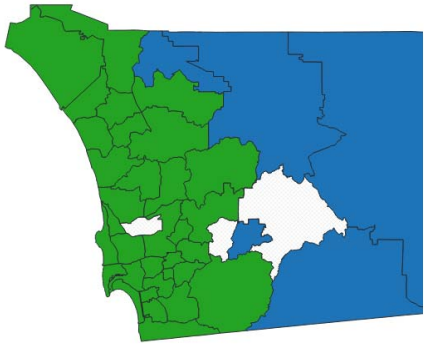

Figure 6. Clustering of subregional areas (SRAs) based on selected neighborhood characteristics

In our study we used this technique to identify groups of regions that behave similarly, either based on health outcomes or on other indicators. Figure 6 shows two clusters of subregional areas (SRAs) that were identified in the San Diego County based on the combination of four neighborhood characteristics: (a) residential access to healthy food sources, (b) residential proximity to fast food locations, (c) park acreage, and (d) sidewalk coverage (all indicators expressed as rates over the population or the size of the SRA). The two clusters are depicted as green and blue, while SRAs that were excluded from the analysis due to insufficient amount of data are shown white. The picture allows researchers that are not familiar with the county to get a quick glimpse in its composition: The western part of the county is more dense, leading to improved accessibility to food sources and increased sidewalk coverage, in contrast to the eastern part of the county, which is more sparsely populated.

Discussion. To conclude, we have seen how machine learning techniques can answer several important public health questions from computing the top determinants of a health outcome, to discovering interesting combinations of indicators that affect the health outcome, to identifying sub-populations that behave in a similar way.

However, while such techniques are useful when public health researchers have very specific questions (e.g., they are interested in the determinants of anxiety disorders), they are less suitable to scenarios where public health researchers want to carry out a more open-ended search for interesting patterns in the data. To enable such scenarios, we augmented the machine learning-driven analytics with a visual data exploration platform, presented next.

B. Visual Data Exploration

In traditional public health studies, researchers typically have a concrete hypothesis they want to check. For instance, they may want to check what is the effect of smoking and sedentary lifestyle on life expectancy [15]. As a result, the study is structured around this hypothesis, and only the specific indicators whose effect on life expectancy they want to test are included in regression modeling that allows them to prove or disprove this hypothesis.

While such hypothesis-driven studies can still be made in the presence of large amounts of data (as explained in the previous section), such big datasets also enable alternate studies that are driven from the data instead. In such cases, researchers start from the data that have been integrated (commonly by a third entity) and try to find interesting patterns that are exposed by the data without trying to prove or disprove a concrete hypothesis. To enable such data-driven studies, we designed and implemented a visual data exploration platform that allows public health researchers to explore public health data by leveraging well-known visualization paradigms. In this Section, we give a brief overview of the tasks that a researcher can carry out using the data exploration platform and describe how it was used in our case study.

Figure 7. Regional Profile for "Pendleton" showing demographics of the selected subregional area

Explore regions. When public health researchers encounter the integrated data for the first time, one of the first tasks they want to carry out is get an overview of the regions that exist in the county and understand their high-level characteristics. To accomplish that, we have created profiles for each region. A regional profile provides a quick overview of the demographics of the region, such as age, race, gender, and income distribution, as well as a list of indicators for which the chosen region has the highest value across all regions.

The regional profiles proved very valuable for our non-local collaborators, as it allowed them to quickly get an overview of the composition of the county. For instance, by looking at the regional profile of the "Pendleton" subregional area (SRA), shown in Figure 7, they could quickly see that the particular SRA has a predominantly young and male population, mostly attributed to the existence of a military base in the area.

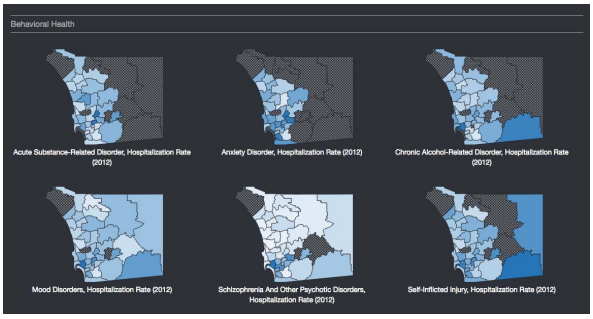

Figure 8. "Small multiples" visualization showing how the hospitalization rates of multiple behavioral health disorders vary across regions

Explore how multiple indicators vary across regions. Once familiar with the regions, the researchers wanted to explore how the regions behaved w.r.t. a selected set of indicators. One of the first requirements was getting an overview of how the values of selected health outcomes varied across regions. It is important to note that due to the exploratory nature of the process, the researchers at this point had not yet narrowed their search down to a particular health outcome. Instead they were focusing on an entire set of health outcomes. For example, in our study on behavioral health outcomes they wanted to see the hospitalization rates for different types of behavioral health conditions, such as acute substance-related disorder, anxiety disorders, chronic alcohol-related disorders, etc. To accommodate this need, we leveraged the "small multiples" visualization paradigm [28]. In particular, we allowed researchers to choose a set of health outcomes and see small maps of how each outcome varied across regions.

Showing indicators on the map, as well as showing all maps next to each other allowed researchers to visually identify spatial patterns that would be hard to infer without a visual interface. For instance, while looking at the "small multiples" visualization of behavioral health outcomes, shown in Figure 8, we made an interesting spatial observation: One of the SRAs has a much smaller hospitalization rate for most behavioral health outcomes, compared to all of its neighboring SRAs that have comparatively high hospitalization rates. Although identifying the reason for this discrepancy is ongoing work, it is a good example of how visualizations can aid in the discovery of interesting patterns that can then be further studied.

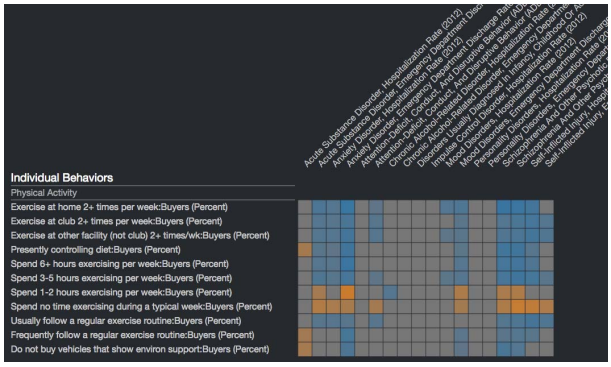

Figure 9. Correlation Matrix showing pairwise correlations between individual behavior indicators and behavioral health outcomes

Explore correlations between indicators. Finally, in an effort to find interesting patterns regarding a set of health outcomes, researchers also wanted to see how different indicators affect a set of health outcomes. To address this need, we implemented a visual correlation matrix that shows pairwise correlations between a set of chosen indicators and a set of selected health outcomes.

In our case study the correlation matrix was used to get a first understanding on how indicators affect multiple health outcome before narrowing down the search to a particular health outcome, which was then studied using machine learning techniques, such as association rule mining. For instance, the correlation matrix shown in Figure 9 was employed to understand how indicators related to individual behaviors relate to health outcomes. Studying the matrix yielded some interesting results: It showed that lower levels of exercising (up to 2 hours per week) are positively correlated with acute substance and

anxiety disorders, while higher levels of exercising are inversely correlated with the aforementioned conditions.

Using the data exploration platform for our case study, we found out that it was a positive first step in allowing our collaborators to quickly gain an understanding of the integrated data and to visually extract interesting patterns that will need to be furthered explored. However, we also identified areas of improvement, which we describe next.

V. Discussion & Future Work

During our case study we explored how one can leverage existing machine learning techniques to answer questions commonly asked on hypothesis-driven epidemiologic studies in the presence of big datasets. Additionally, we studied how data visualization techniques can be used to enable big data-driven epidemiologic studies. Both approaches were well-received by our public health collaborators, who recognized the usefulness of these techniques. However, in the process, we also identified several challenges that still need to be addressed. We next describe these challenges, which we are planning to address as part of our future work.

Information overload. While machine learning and data visualization techniques help researchers quickly identify interesting patterns from a very large dataset (which would be unmanageable without the use of such techniques), these techniques often produce significant amount of information that requires significant time from the users to digest. For instance, feature selection yields a large number of indicators that potentially determine life expectancy, many of which are already known in the literature (e.g., all socioeconomic factors). Similarly, the correlation matrix can be overwhelming when studying pairwise correlations between large sets of indicators and health outcomes.

To reduce the output of the above techniques to more manageable levels, we are going to utilize a combination of techniques: First, we will incorporate domain knowledge into the algorithms, so that they avoid generating known results. We are currently experimenting with extracting domain knowledge from publicly available ontologies, as well as from annotations placed by the users on the data. Second, we will extend the algorithms to create high-level summarizations (i.e., clusters) of the results, so that the user can quickly get a quick overview of the results and then drill-down into specific areas of interest. Our preliminary experiments on clustering the indicators shown on the correlation matrix into groups of indicators with similar correlation patterns indicate that this is a promising direction for future work.

From interesting patterns to discoveries. Another observation we made during our case study is that, even though the analytics considered above lead to the discovery of interesting associations, it is not clear whether these associations are important or whether they are the result of other latent associations. For instance, is the existence of parks really related to life expectancy or is the apparent association caused by underlying socioeconomic factors? The theoretical toolbox of a public health researcher contains techniques for answering such questions. In particular, when studying associations between indicators and health outcomes, researchers commonly control the associations for known confounders (i.e., other indicators that relate to the health outcome and may partially explain the association). To allow researchers to answer such questions and go from seemingly interesting patterns to discovery, we will extend existing machine learning algorithms (such as association rule mining) with common tasks used in public health research (such as controlling for confounders).

Spurious discoveries. Last but not least, an issue that has to be addressed is spurious discoveries. It is well known that by increasing the number of associations tested, one also increases the probability of finding spurious associations. Existing works try to alert the user of this possibility by adjusting the importance of the discovered associations based on the number of tests, using the Bonferroni correction [4] or other metrics [3]. While such approaches are a step in the right direction, they may also lead to scenarios where all associations are marked as non-important due to the use of corrections that are shown to be very conservative. To address this problem, we are working towards limiting the number of tests carried out in the first place by allowing the user to guide the search. This can be easily combined with the clustering approach mentioned earlier in the section, according to which the system produces first high-level results to help the user guide the search into the area of her interest.

We are currently working to address the above issues by creating a novel visual and interactive "epidemiologist's workbench", which will allow public health researchers to efficiently analyze large amounts of data and gain interesting insights into the determinants of health conditions. The workbench will combine both the machine learning techniques presented in Section IV.A and the data exploration paradigms presented in Section IV.B into a common visual interface, allowing public health researchers to seamlessly transition from data-driven exploration to hypothesis-driven analysis. Given the promising results from our preliminary case study, we believe that the resulting system will be an important tool for public health researchers, that will help them make interesting discoveries from the continuously increasing amounts of health-related data.

Acknowledgement

The authors would like to thank the team members of the collaborating agencies that participated in this study from the Center on Society and Health at the Virginia Commonwealth University (Amber Haley, Kristin Smith, Lauren Snellings, and Emily Zimmerman) and the San

Diego County Health and Human Services Agency (Nick Macchione, Dale Fleming, and Leslie Ray). We also thank our funders, the National Science Foundation (NSF/IIS 1237174) and the Robert Wood Johnson Foundation (72382), who supported this work.

REFERENCES

[1] Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases." In Acm sigmod record, vol. 22, no. 2, pp. 207-216. ACM, 1993.

[2] Bailey, S. and Handu, D., 2012. Introduction to Epidemiologic Research Methods in Public Health Practice. Jones & Bartlett Publishers.

[3] Binnig, Carsten, Lorenzo De Stefani, Tim Kraska, Eli Upfal, Emanuel Zgraggen, and Zheguang Zhao. "Towards Sustainable Insights, or Why Polygamy is Bad for You" Conference on Innovative Data Systems Research (2017).

[4] Bonferroni, Carlo E. Teoria statistica delle classi e calcolo delle probabilita. Libreria internazionale Seeber, 1936.

[5] Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. "Classification and regression trees. Wadsworth & Brooks." Monterey, CA (1984).

[6] Breiman, Leo. "Bagging predictors." Machine learning 24, no. 2 (1996): 123-140.

[7] Dadvand, Payam, Bart Ostro, Francesc Figueras, Maria Foraster, Xavier Basagaña, Antònia Valentín, David Martinez et al. "Residential proximity to major roads and term low birth weight: the roles of air pollution, heat, noise, and road-adjacent trees." Epidemiology 25, no. 4 (2014): 518-525.

[8] Doll, Richard, and A. Bradford Hill. "Smoking and carcinoma of the lung." British medical journal 2, no. 4682 (1950): 739.

[9] Fang, Ruogu, Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, and S. S. Iyengar. "Computational health informatics in the big data age: a survey." ACM Computing Surveys (CSUR) 49, no. 1 (2016): 12.

[10] Friis, Robert H., and Thomas Sellers. Epidemiology for public health practice. Jones & Bartlett Publishers, 2013.

[11] Institute of Medicine (US). Committee for the Study of the Future of Public Health. The future of public health. Vol. 88, no. 2. National Academy Press, 1988.

[12] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31, no. 3 (1999): 264-323.

[13] Kaufman, Leonard, and Peter Rousseeuw. Clustering by means of medoids. North-Holland, 1987.

[14] Khoury, Muin J., and John PA Ioannidis. "Big data meets public health." Science 346, no. 6213 (2014): 1054-1055.

[15] Kvaavik, Elisabeth, G. David Batty, Giske Ursin, Rachel Huxley, and Catharine R. Gale. "Influence of individual and combined health behaviors on total and cause-specific mortality in men and women: the United Kingdom health and lifestyle survey." Archives of internal medicine 170, no. 8 (2010): 711-718.

[16] Malloy, Michael H., and Daniel H. Freeman. "Birth weight- and gestational age-specific sudden infant death syndrome mortality: United States, 1991 versus 1995." Pediatrics 105, no. 6 (2000): 1227-1231.

[17] McKeown, Thomas. "The role of medicine: dream, mirage or nemesis? 1979." London: Nuffield Provincial Hospitals Trust 180.

[18] Piatetsky-Shapiro, Gregory. "Discovery, analysis, and presentation of strong rules." Knowledge discovery in databases (1991): 229-238.

[19] Quinlan, J. Ross. "Simplifying decision trees." International journal of man-machine studies 27, no. 3 (1987): 221-234.

[20] Raghupathi, Wullianallur, and Viju Raghupathi. "Big data analytics in healthcare: promise and potential." Health Information Science and Systems 2, no. 1 (2014): 3.

[21] Rothman, Kenneth J., Sander Greenland, and Timothy L. Lash, eds. Modern epidemiology. Lippincott Williams & Wilkins, 2008.

[22] Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics 20 (1987): 53-65.

[23] Santillana, Mauricio, André T. Nguyen, Mark Dredze, Michael J. Paul, Elaine O. Nsoesie, and John S. Brownstein. "Combining search, social media, and traditional data sources to improve influenza surveillance." PLoS Comput Biol 11, no. 10 (2015): e1004513.

[24] Savasere, Ashok, Edward Robert Omiecinski, and Shamkant B. Navathe. An efficient algorithm for mining association rules in large databases. Georgia Institute of Technology, 1995.

[25] Sedig, Kamran, and Oluwakemi Ola. "The challenge of big data in public health: an opportunity for visual analytics." Online journal of public health informatics 5, no. 3 (2014).

[26] Solar, Orielle, and Alec Irwin. "A conceptual framework for action on the social determinants of health." (2007).

[27] Tan, Pang‐Ning, Michael Steinbach, and Vipin Kumar. "Data mining cluster analysis: basic concepts and algorithms." Introduction to data mining (2013).

[28] Tufte, Edward R. "The visual display of quantitative information." Graphics Press (2001).

[29] Woolf, S. H., and L. Aron, eds. "U.S. Health in International Perspective: Shorter Lives, Poorer Health." Panel on Understanding Cross-National Health Differences Among High-Income Countries. National Research Council, Committee on Population, Division of Behavioral and Social Sciences and Education, and Board on Population Health and Public Health Practice, Institute of Medicine. Washington, DC: The National Academies Press, 2013.

[30] Wynder, Emst L. "Invited commentary: Studies in mechanism and prevention: Striking a proper balance." American journal of epidemiology 139, no. 6 (1994): 547-549.