Assisting Discovery in Public Health *

Yannis Katsis† ikatsis@cs.ucsd.edu
Nikos Koulouris† nkoulour@cs.ucsd.edu
Yannis Papakonstantinou† yannis@cs.ucsd.edu
Kevin Patrick‡ kpatrick@ucsd.edu

†Dept. of Computer Science and Engin. ‡Dept. of Family Medicine and Public Health & The Qualcomm Institute
University of California, San Diego

ABSTRACT

Several public health (PH) researchers have lately been arguing that big data can play a profound role in scientific discovery. Leveraging the vast amount of population-level data collected by public agencies and other organizations, could lead to important discoveries that were not necessarily suspected to be true. However, they also warn about the pitfalls of data-driven discovery: The large amount of data can easily lead to information overload for the researchers. Additionally, data-driven studies that make a lot of tests in the search for important discoveries have the potential to lead to discoveries that seem important but are in fact random.

We show that data-driven studies can be effective and yet avoid the potential pitfalls by keeping the researchers in the loop of the discovery process. To this end, we propose PHD; an interactive visual discovery system that allows public health researchers to gain interesting insights from large datasets. PHD generalizes the current workflow of PH researchers by facilitating the major analytics tasks involved in PH discovery, such as calculating important associations based on the standard notions of odds rations and confidence intervals, controlling for the effect of other variables and discovering interesting compounding effects. More importantly however, it leverages user interaction and the semantics of the domain to make sure that this workflow scales to large datasets, while avoiding information overload and random discoveries.

1. INTRODUCTION

Public health is the science of preventing disease and promoting health and well-being. To accomplish this goal, public health researchers (also known as epidemiologists) aim in discovering the determinants of health; i.e., in the factors that affect health outcomes. For instance, they may be interested in the factors that are responsible for low weight of infants at birth. Such discoveries are made through public health studies.

Traditionally, public health studies are *hypothesis-driven*. Epidemiologists start from a hypothesis about a (typically limited) set of factors that are strongly suspected to affect an outcome. They then collect data about these factors and outcome and check whether their hypothesis holds. For example, the authors of [15] suspected that there is a correlation between maternal residential proximity to major roads and low birth weight. In order to test this hypothesis, they collected data about the distance from major roads and other related measures, such as air pollution and noise, and used them to test their hypothesis.

Figure 1: Screenshot of PHD's discovery explorer

Starting from hypotheses though makes it hard to make unexpected discoveries about factors that are not even suspected to influence a health outcome. To address this concern, many public health experts are pushing for a *data-driven* approach to public health. They suggest that by leveraging large volumes of data that are collected about individuals and/or communities (including hospitalizations, socioeconomic data, behavioral data, and environmental data), we may be able to make unexpected and highly valuable discoveries about the determinants of health [8, 10, 11]. For instance, imagine leveraging a large dataset on environmental data to study birth weight. Using this dataset, one could produce evidence not only for the association of the birth weight with major roads, but also for the association with other environmental factors that one may have never expected (e.g., proximity to parks).

However, data-driven public health studies do not come without pitfalls. While hypothesis-driven studies are focused analyses that involve a limited set of factors, data-driven studies start from a considerably larger pool of factors (in the order of hundreds or thousands). As such, data-driven studies are hard to carry out without extra software support, as the current workflow of epidemiologists does not scale to large numbers of factors: It is hard not only to run their analyses for a large number of factors using their tools of choice (typically spreadsheets or statistical packages), but also to gain an overview of the multitude of results that would arise from

*This work was supported by NSF IIS-1237174.

HILDA'17, May 14, 2017, Chicago, IL, USA

DOI: http://dx.doi.org/10.1145/3077257.3077269

	Positive Outcome h	Negative Outcome $\neg h$
Exposed f	P_E	N_E
Not Exposed $\neg f$	P_N	N_N

Table 1: Contingency table for $\langle f \rightarrow h \rangle$

such a process. Last but not least, by testing for a large number of associations, they are very likely to make spurious discoveries. As many public health experts warn, increasing the number of tests also increases the probability of discoveries that seem important, but are in fact random [8].

We argue that data-driven studies that simultaneously have the potential for important, unexpected discoveries and avoid potential pitfalls regarding lack of overview and spurious discoveries are possible by keeping the epidemiologist in the loop during the discovery process. To this end, we present *PHD (Public Health Discovery)*; an interactive visual platform that allows epidemiologists to guide the discovery process and quickly make interesting discoveries in large datasets. PHD not only supports the tasks commonly carried out by researchers in traditional hypothesis-driven studies (such as finding the most interesting discoveries based on the same formal statistical guarantees used in the field), but also makes sure that these tasks scale to large data sets. Finally, it exploits the user interaction and the semantics of the data to reduce the probability of false discoveries. Figure 1 depicts PHD's main screen, which we describe in detail in later sections. This work is based on observations we made while integrating and analyzing data on more than 3,800 combined factors and outcomes about regions of San Diego County, in the context of UC San Diego's DELPHI project [7].

Contributions. This work makes the following contributions:

- It describes and formalizes the workflow followed by epidemiologists in traditional hypothesis-driven studies.
- It identifies the challenges of adjusting this workflow to data-driven scenarios, which typically include a large number of variables, making the problem substantially different from the traditional small-data studies.
- It describes a novel visual interactive platform that enables a new workflow suitable for data-driven studies. The platform supports the notions and statistical guarantees commonly accepted in public health, while adjusting them for the big amounts of data found in data-driven use cases.
- It discusses how to leverage user interaction and domain-specific semantics to address the problem of false discoveries, commonly found in data-driven exploration systems.

2. DISCOVERIES IN PUBLIC HEALTH

Before we explain how to facilitate data-driven epidemiological studies, we first describe the workflow followed by epidemiologists in current hypothesis-driven studies. Researchers conducting a study typically start from a hypothesis that certain *health factors* (in short *factors*), such as proximity to major roads and air pollution, are correlated with a particular *health outcome* (in short *outcome*), such as low birthweight. The goal of the study is then to prove that one or more of these factors f have a strong association to the health outcome h. For the purposes of this work, this association is called a *discovery*, and denoted by $\langle f \rightarrow h \rangle$. However, not all discoveries are interesting. For a discovery to be interesting it has to be statistically significant. Thus the first step in an epidemiologist's workflow is to find statistically significant discoveries.

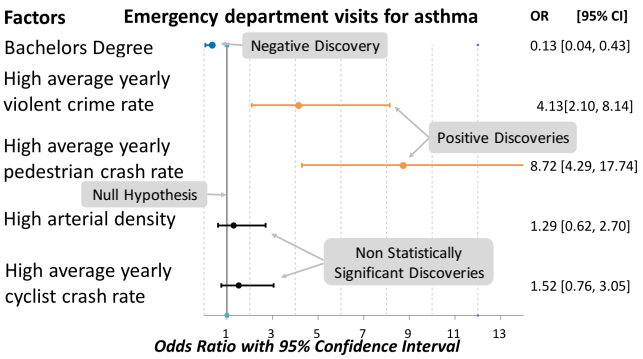


Figure 2: Epidemiologist's worksheet showing the association of five factors with asthma emergency department visits

Finding statistically significant discoveries. Statistics propose a multitude of metrics to measure statistical significance. In public health, the commonly accepted standard is the odds ratio and the associated confidence interval. The *odds ratio (OR)* of a discovery $\langle f \rightarrow h \rangle$ is the ratio of the odds of having a health outcome h while being exposed to a factor f (i.e., proximity to major roads), divided by the odds of having the outcome and not being exposed to the factor. The *confidence interval (CI)* is used to measure the statistical significance of the discovery. The medical field, typically relies on 95% CI. Formally, the OR and the 95% CI of a discovery are computed as follows:[1]

DEFINITION 2.1. Odds Ratio & 95% Confidence Interval. *Given a discovery $\langle f \rightarrow h \rangle$, consider its contingency table shown in Table 1. Each cell of the table denotes the number of subjects that satisfy the variables of the corresponding row and column.The odds ratio of the discovery is given by:*

$$OR(\langle f \rightarrow h \rangle) = \frac{\frac{P_E}{N_E}}{\frac{P_N}{N_N}} = \frac{P_E N_N}{N_E P_N}$$

while its 95% confidence interval is the range $[lower\ limit\ (ll), upper\ limit\ (ul)]$, whose upper and lower limits are given by:

$$[ll, ul] = e^{[ln(OR) \pm 1.96\sqrt{1/P_E + 1/N_E + 1/P_N + 1/N_N}]}$$

To compare discoveries, epidemiologists typically employ statistical packages to visualize the odds ratios and confidence intervals through forest plots. Figure 2 shows a forest plot of the association of five different factors with asthma emergency department visits (the data shown are extracted from the DELPHI dataset). An odds ratio of 1 (a condition known as *null hypothesis*), signifies that the factor does not have any effect on the health outcome. Therefore, a discovery is considered statistically significant if its confidence interval does not contain the value 1. In Figure 2, only three out of the five discoveries are statistically significant.

Controlling for known confounders. Finding statistically significant discoveries is only the first step. A statistically significant discovery could still be non-interesting if it is caused by a third factor that is known to be associated with the health outcome. For instance, a discovery that the birthweight is associated with proximity

[1]For a complete review of odds ratios and confidence intervals, the reader is referred to [9].

to major roads may be irrelevant if it is caused by socioeconomic factors (e.g, income) and these factors are known determinants of low birth weights. Such factors that are associated with both the premise and the conclusion of a discovery are called *confounders*. The public health literature contains extensive lists of factors that are known to affect certain health outcomes and are therefore used as confounders, with socioeconomic factors being the most prominent example. To discard discoveries caused by known confounding factors, as a second step in the analysis, epidemiologists *control* for such factors, by adjusting the odds ratio, so that the effect of the confounders is removed. The odds ratio of a discovery controlled for a factor is computed as follows:

DEFINITION 2.2. Controlling for a factor. *Let $\langle f \rightarrow h \rangle$ be a discovery and c a factor to control for. Let the contingency table of the subpopulation that satisfies c (resp. $\neg c$) be Table 1 with the suffix $'|c'$ (resp. $'|\neg c'$) added to all variables. Then the odds ratio of the discovery controlled for c is given by:*

$$OR^{c}_{adj}(\langle f \rightarrow h \rangle) = \frac{\frac{P_{E|c}N_{N|c}}{|c|} + \frac{P_{E|\neg c}N_{N|\neg c}}{|\neg c|}}{\frac{N_{E|c}P_{N|c}}{|c|} + \frac{N_{E|\neg c}P_{N|\neg c}}{|\neg c|}}$$

where $|c|$ (resp. $|\neg c|$) is the count of subjects satisfying c (resp. $\neg c$)

Finding interesting compounders. Once epidemiologists find an interesting discovery (i.e., one that is statistically significant even after controlling for known confounders), they next study how the discovery is affected by other factors. For example, does proximity to parks affect the premise of the initial discovery? Such factors, which when added to the premise of the discovery significantly affect its statistical significance (by either increasing it if they have a positive effect or decreasing it if they have a negative effect), are called *compounders*.

Epidemiologists want thus to find the factors $f_1, f_2, \ldots, f_n$ which when added to a discovery $d : \langle f \rightarrow h \rangle$ yield a new discovery $d' : \langle f, f_1, f_2, \ldots, f_n \rightarrow h \rangle$, which is still statistically significant and whose odds ratio $OR(d')$ is significantly different from the odds ratio $OR(d)$ of the initial discovery. Note that one may argue that by adding more factors to the discovery, it is easy to achieve a high odds ratio by limiting the population that satisfies the discovery (i.e., minimizing P_N). This is where the importance of the confidence interval shines, as these discoveries will easily stop being statistically significant.

At the end of this discovery process, epidemiologists have identified a set of discoveries with one or more factors in their premise that are both statistically significant and not explained by known confounders.

3. PHD: THE DISCOVERY SYSTEM

In contrast to hypothesis-driven studies that focus on the interaction between a limited set of factors and outcomes, data-driven approaches promise the analysis of hundreds or thousands of factors and outcomes. Typically, these are data sets that have been created by accumulating population-level data from a variety of sources, including health agencies, hospitals, surveys, environmental reports, etc. This explosion in the number of factors and outcomes explored creates significant challenges:

- *Lack of overview:* Due to the large number of factors, it is hard for the epidemiologist to have an overview of all the data that are used in the analysis.
- *Information overload:* Similarly, the large number of potential discoveries have the potential of overwhelming the user if they are not appropriately ranked or summarized.
- *Random discoveries:* Finally, due to the many associations tested, it is easy to arrive at seemingly interesting discoveries that are random.

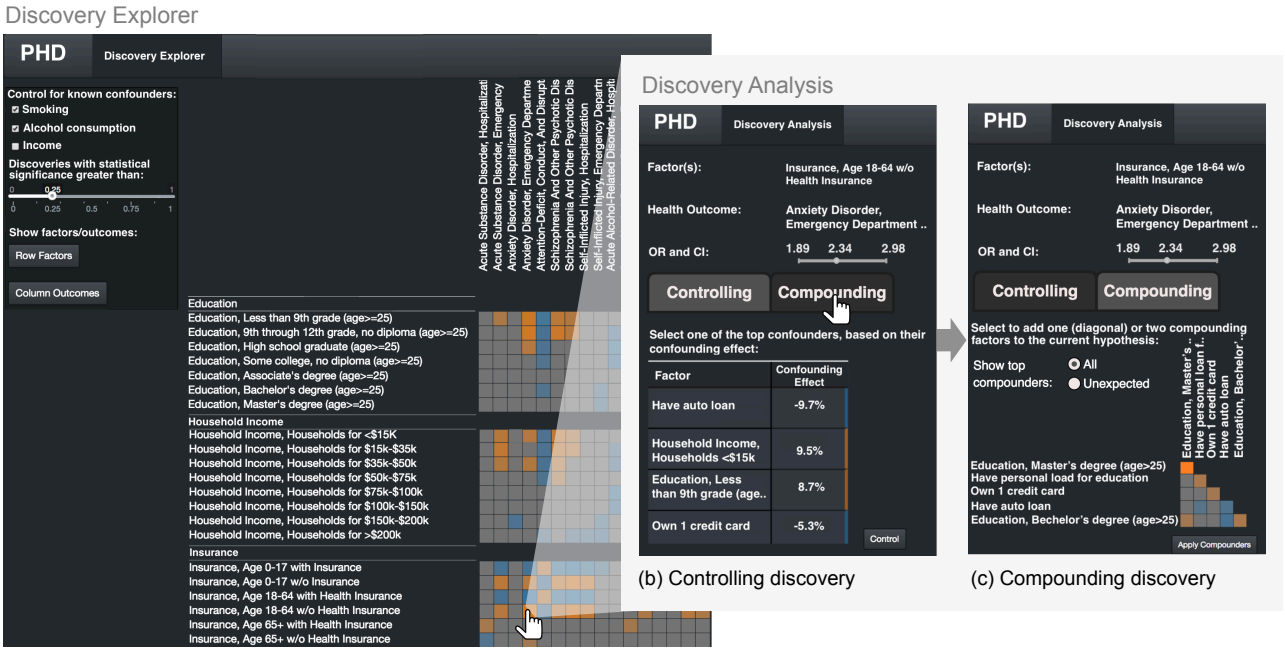
We next describe how PHD addresses these challenges.

Running example. We use as our running example a data-driven study carried out at UC San Diego in collaboration with the Center on Society and Health at the Virginia Commonwealth University and the San Diego Health and Human Services Agency. As part of the study, we integrated data about more than 3,800 combined factors and outcomes, including environmental exposures (such as traffic density and air pollution), individual behaviors (such as smoking, exercising, and consumer buying patterns), health systems (such as insurance status), and hospitalizations and emergency department visits for various health conditions. We then analyzed the data using a preliminary version of the proposed system to get insights on the determinants of *behavioral health conditions* (such as anxiety disorders, attention-deficit disorders, schizophrenia, etc.). A demo of this early version of PHD can be found in [2]. The lessons learned from the study were used to inform the revision of the system, which led to the system presented below.

We next outline how PHD facilitates the three main steps of the discovery process: (a) finding statistically significant discoveries, (b) controlling for confounders, and (c) exploring compounders.

Finding statistically significant discoveries. As a first step in the discovery process, PHD allows epidemiologists to find statistically significant discoveries. However, the plethora of factors and outcomes makes it impossible to present to the user all discoveries whose confidence interval does not contain the null hypothesis (which is the definition of statistical significance, as explained in Section 2). To avoid information overload, PHD limits the sets of discoveries displayed on the screen in two ways:

- *User guidance:* The epidemiologist guides the search by selecting a subset of factors and outcomes of interest. Even though it may at first sight seem counter to the data-driven nature of the system, in reality epidemiologists are not interested in a blind search for discoveries. Instead, they have some broad area of interest in mind. For instance, in our running example, the user may be interested in exploring how factors related to education, insurance, and income affect behavioral health outcomes. To facilitate this user guidance, all factors and outcomes are grouped into broad areas and subareas (such as socioeconomic factors, individual behaviors, etc.).
- *Top k computation:* However, even with the user's restriction on the input, the number of statistically significant discoveries may still be substantial. To address this problem, PHD computes and displays the top k discoveries, in terms of significance that contain factors and outcomes selected by the user (where the value of k is either selected by the user or automatically chosen by the system based on the available screen real-estate to reduce the need for scrolling). An interesting problem is selecting the metric that is used for the top k computation. As described in Section 2, the significance of a discovery is represented by its confidence interval, which is described by two numbers (i.e., the two interval limits). In traditional studies, epidemiologists could visualize and display the limited number of confidence intervals so that they could compare them, as shown in Figure 2. In data-driven studies, the number of discoveries precludes such an approach. Instead, PHD represents the significance of a discovery through a single *normalized significance value* that is used for the top k computation and which is given by the following formula:



(a) Exploring determinants of behavioral health outcomes

(b) Controlling discovery

(c) Compounding discovery

Figure 3: Example interaction of epidemiologist with PHD

$$\text{Normalized Significance (NS)} = \begin{cases} \frac{1-ul}{ul+1}, & \text{negative discovery} \\ \frac{ll-1}{ll+1}, & \text{positive discovery} \end{cases}$$

This value, corresponds to the limit of the confidence interval that is closest to the null hypothesis (which is intuitively the worst-case odds ratio). Additionally, this value normalizes positive and negative discoveries (as the range of significance for negative and positive discoveries is [0,1] and [1, $+\infty$], resp.).

PHD visualizes the resulting k discoveries through the *discovery explorer* shown in Figure 3a. Rows and columns correspond to factors and health outcomes, resp., participating in the top k discoveries. Similar factors are grouped into large semantic categories (e.g., education or insurance), which exist in the input data, as we will discuss in Section 4. Each discovery $\langle f \rightarrow h \rangle$ is shown as a colored cell at the intersection of the row/column corresponding to the factor f/outcome h, respectively. A blue (resp., orange) shade denotes negative (resp., positive) discoveries, while the color intensity represents the normalized significance of the discovery (more saturated colors denote more significant discoveries). For instance, the discovery under the mouse pointer denotes a positive association between lack of health insurance for 18-64 year olds and high emergency department discharge rates for anxiety disorders.

Controlling for known (and unknown) confounders. At any point during the discovery process, the epidemiologist can control the discoveries for confounders. To ease the process, PHD incorporates knowledge about known confounders for particular health outcomes. Known confounders for the outcomes shown on the screen are shown on the left-hand side of the discovery explorer for the epidemiologist to select. Upon selection, PHD controls for these factors and updates the top-k discoveries and their significance values. For instance, in Figure 3a, the user has already controlled for smoking and alcohol consumption.

In addition to being semantically important, controlling for known factors is also very beneficial in data-driven scenarios for quickly pruning non-interesting discoveries. Preliminary results show that controlling can considerably reduce the number of discoveries to be explored. For instance, in an experiment with 810 factors and 1 health outcome, controlling for households that received food stamps (a proxy for income) reduced the number of statistically significant discoveries from 362 to 203 (a reduction of 44%).

However, not all confounders are known. A discovery may not have any known confounders, but it may still have confounders that were not previously known. Intuitively, these correspond to unknown latent variables that explain a discovery. In hypothesis-driven studies, finding unknown confounders is rare, as only a limited set of factors, hand-picked by the researchers, participate in the study. In data-driven scenarios on the other hand, the large set of participating factors significantly increases the potential for finding interesting unknown confounders.

This is why PHD, in addition to known confounders, allows epidemiologists to find and control for unknown confounders. Once the user selects a discovery from the discovery explorer, PHD shows the *discovery analysis* screen (see Figure 3b). This screen, among others, shows the top unknown confounders for the selected discovery. A confounder c of a discovery d is ranked based on its *confounding effect*, which is defined as $|\frac{NS(d)-NS^c_{adj}(d)}{NS(d)}|$, where $NS(d)$ and $NS^c_{adj}(d)$ is the normalized significance value of discovery d before and after controlling for confounder c, respectively. Intuitively, the confounding effect of c expresses the relative change of the discovery's significance value after controlling for c. After reviewing the top unknown confounders, an epidemiologist can control for any of them, by simply selecting them. For instance, Figure 3b shows the top unknown confounders for the association between the lack of health insurance for 18-64 year olds and anxiety disorder emergency department discharge rates. The screen shows that if we control for subjects that have auto loans, the normalized significance of the discovery is reduced by 9.7%. Note that the number of such unknown confounders could be significant. While by showing the top confounders we avoid information overload, in our future work we will explore how we can leverage the

semantics of the data (which as we will see in Section 4 are important for other purposes as well) to select the unknown confounders that are more likely to be of interest to the user.

Discovering interesting compounders. Once an epidemiologist has found an interesting discovery, they can use the compounding tab of the discovery analysis screen to also check for interesting compounding factors (see Figure 3c). The tab shows the top compounding factors for the selected discovery d, ranked by their *compounding effect*; i.e. the relative change to the normalized significance of d when compounding factor c is added to d's premise. The top compounding factors are shown on an array, which allows the user to select one factor (by selecting a cell on the diagonal) or two factors (by selecting any other cell), to be added to the discovery. Once the user makes a selection, the chosen factor(s) are added to the discovery and its confidence interval is updated. For instance, Figure 3c shows that having a Master's degree acts as a strong compounder for the association between lack of health insurance for 18-64 year olds and anxiety disorder emergency department visits.

PHD faciltates the discovery of two types of compounding factors, pruning along the way across myriads of combinations. In addition to the commonly studied compounding factors (which as explained in Section 2 are defined as factors that significantly affect a discovery when added to its premise), PHD allows the user to also find *unexpected compounding factors*. We call a factor f_1 an unexpected compounding factor for a discovery $\langle f \rightarrow h \rangle$ iff $NS(\langle f, f_1 \rightarrow h \rangle) \gg$ (or $\ll$) $NS(\langle f \rightarrow h \rangle) * NS(\langle f_1 \rightarrow h \rangle)$; i.e., iff the joint effect of f_1 and f on h is significantly different than the multiplication of their individual effects. This intuitively shows that f and f_1 are not independent factors, but have an interesting interaction w.r.t. the health outcome h. This is an important feature that enables discoveries, such as the one in [4], which states that obesity and smoking are both risk factors for cardiovascular diseases, but together they have an amplified effect. The discovery analysis screen allows a user to choose between the top compounders in general and the top unexpected compounders to facilitate different types of discoveries.

4. AVOIDING RANDOM DISCOVERIES

While we have shown how PHD addresses the lack of overview and information overload pitfalls inherent in data-driven studies, we have not yet seen how it deals with random discoveries. As the critics of data-driven approaches correctly argue, as a system tests more and more associations in order to arrive at a significant discovery, the probability of finding a discovery that appears statistically significant but is in fact random increases as well [8]. This phenomenon, well known in both statistics and medicine [5], is particularly pronounced in data exploration systems, which make it easy to perform large numbers of tests without even realizing it [6].

To mitigate this problem, statisticians have proposed a variety of methods that adjust upwards the probability associated with the confidence interval (upwards of the conventional 95%) to account for the fact that one has computed multiple associations. Consequently, the confidence intervals of individual discoveries widen. These include the well-known Bonferonni correction (which is considered too conservative), the Benjamini-Hochberg procedure [5], and others. Recently, [6] integrated such metrics in a visual exploration system.

Limiting the number of tests. While important steps in the right direction, these approaches diagnose and mitigate the problem *after the fact* (i.e., after a large number of associations have been tested, which may have led to very wide confidence intervals, which in turn

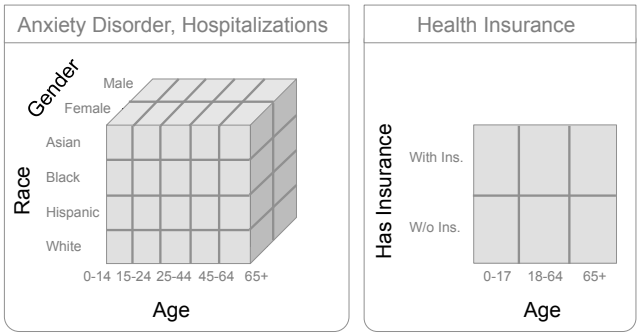


Figure 4: Semantic structure of factors and outcomes

may have led to false negatives where legitimate discoveries are deemed non-significant ones). Interactive systems, such as PHD, offer the opportunity to combine this with a *proactive* approach that leverages user interaction to avoid performing many tests in the first place. By receiving guidance from the user on which discoveries to pursue, the system can avoid testing and paying the penalty in statistical significance for associations of no interest to the user.

At first sight, this seems to go against the data-driven nature of the system. How can a user inform the system that they are not interested in a discovery before knowing the statistical significance of said discovery? Ontologies come to the rescue at this point: To solve this apparent circularity, the system groups factors into higher-level categories and first displays discoveries in terms of these high-level categories. For the purpose of the top-k discovery computation each category acts as a single factor and thus computing the association of the category with an outcome requires significantly fewer tests than testing the associations of all factors within the category with the outcome. After seeing the discoveries in terms of categories, the user can select a category of interest and drill further down into it.

To group factors into categories, PHD exploits the semantic structure of factors and outcomes that typically exists in the source data (e.g., in the ontology implied by the hierarchical structure of questionnaires or in explicit medical ontologies [1]). In particular, the input data typically contain characteristics of the population that are broken down into finer-grained factors, according to a set of dimensions. For instance, Figure 4 shows two such broad characteristics; health insurance and hospitalizations for anxiety disorders. The latter is broken down according to gender, race, and age, while the former is broken down according to health insurance status and age. If flattened, each characteristic yields as many factors as the combination of values for its dimensions. It is therefore obvious that using this structure to cluster the flat factors into higher-level categories (e.g., total of people with/without health insurance regardless of age), can significantly cut down on the number of tests performed by the system.[2]

Figure 1 on the first page of the paper shows how the discovery explorer of Figure 3a is transformed by the use of categories that exploit the semantic structure of the data (and after drilling down into the non-insured population category). For instance, the six health insurance factors of Figure 3a have been summarized through two high-level factors (with and without health insurance).

[2]Note that in contrast to data warehouses where data are commonly modelled as a single datacube, public health data are clustered into *multiple* datacubes. The reason is that they are population-level data that do not contain certain combinations of dimensions. For instance, even though we know how many people were hospitalized for anxiety disorders and how many have health insurance, we may not know how many people simultaneously satisfy both conditions.

With grouping, not only becomes the visualization more concise, but the system also performs significantly fewer tests, thus reducing the possibility of random discoveries.

Trading off statistical significance with information loss. However, summarizing discoveries into higher-level categories comes with an information loss, as discoveries that exist at lower levels may not appear at the higher level. For instance, consider the non-insured population category shown in Figure 1 and its children. Most of the age groups of uninsured subjects behave similarly and hence they are correctly summarized by the uninsured category. Uninsured over 65 years though behave differently than other age groups by having a negative association with hospitalizations for anxiety disorders. While this discovery may be interesting, it does not appear on the discovery explorer if we bundle together all uninsured into a single category. Each grouping of factors into categories corresponds thus to a particular tradeoff between statistical significance and information loss, with grouping into broader categories exhibiting higher statistical significance but also increased information loss. Currently, PHD groups factors into categories that are predetermined by domain experts (e.g., the expert might decide that it makes more sense to group health insurance by the status and not by the age). Using the data to automatically select the grouping that corresponds to the optimal tradeoff is though an interesting problem that we will pursue in our future work.

5. RELATED WORK

Leveraging big data to discover interesting patterns has been the focus of extensive work in three areas: public health, data mining, and databases. In addition, the statistics community has done a lot of work on reducing the possibility of false discoveries.

The *public health community* has lately started using data-driven techniques to compute factors that affect an outcome [10]. However, such approaches are limited to particular studies and are not concerned with automating the process. Even though it has been argued that public health has a lot to gain from big data [8, 11], we are not aware of specialized interactive discovery tools for the problem, which is the focus of our work.

The *data mining and machine learning community* has worked on inferring associations from a dataset, commonly known as association rule mining (ARM) [3]. However, ARM algorithms are typically based on non-standard statistical notions (commonly confidence and support). The limited work that exists on ARM with odds ratios does not support confidence intervals, which form the accepted standard for statistical significance in public health. Second, PHD can prune the space of compound effects (which are equivalent to associations in ARM) by dismissing those that are implied by the independence assumption of their components. Finally, our work helps with the summarization of discoveries. Note that the latter two aspects address the most common complaint about ARM techniques, which is the exorbitant size of their output.

The *database community* has lately worked on general frameworks for extracting and visualizing interesting facts. In particular, systems such as SEEDB [14] and zenvisage [13], have focused on extracting automatically or through a specification language, visualizations that exhibit patterns of interest to the user. Compared to these works which look at general notions of "interestingness", our work focuses on interesting discoveries and provides a discovery process tailored to public health. To achieve that, our work augments visual discovery with the statistical guarantees needed in the field, it aims to help with the whole workflow of a researcher by facilitating pruning spurious discoveries at different stages of the discovery process, and it places significant weight on user interaction as a way of improving the provided guarantees.

Related to the summarization of the results done by PHD is work on clustering the data as a way of providing an overview in the exploratory process [12]. However, in contrast to our work, the summarization uses the actual data (and not their semantics) and, more importantly, the summarization is used to improve user interaction (not any associated statistical guarantees, as is the case with PHD).

Finally, the *statistics community* has proposed several techniques to solve the problem of maintaining the family-wise error rate (i.e., adjusting the statistical significance of associations for past tests). These include the Bonferroni correction, the Benjamini-Hochberg method [5], and many more. The reader is referred to [6] and [16] for a discussion and detailed evaluation of the effectiveness of such techniques for interactive data exploration.

6. CONCLUSION & FUTURE WORK

Public health can greatly benefit from augmenting the common hypothesis-driven studies with data-driven studies that leverage big amounts of health-related data to find the determinants of health. In this work, we have shown how we can leverage user interaction and the semantics of the domain to facilitate data-driven studies that lead to interesting discoveries and avoid common pitfalls, such as the information overload and the potential for random discoveries.

Future work will explore further the semantics-based approach. In particular, semantics will be used (a) to augment the top k computation of confounders and compounders, in order to present the confounding and compounding factors that are more likely to be of interest to the user, and (b) to find the grouping of factors and outcomes that corresponds to the right tradeoff between statistical significance and information loss, as explained in Section 4. Once we have solved the technical problems, we will be evaluating the PHD platform with public health experts in the DELPHI project.

7. REFERENCES

[1] SNOMED CT, http://www.snomed.org/snomed-ct.

[2] Video of early prototype of PHD, https://youtu.be/vmzguym98w0?t=25m00s.

[3] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD record*, 1993.

[4] M. Akbartabartoori, M. E. Lean, and C. R. Hankey. Smoking combined with overweight or obesity markedly elevates cardiovascular risk factors. *European Journal of Cardiovascular Prevention & Rehabilitation*, 13(6):938–946, 2006.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 1995.

[6] C. Binnig, L. D. Stefani, T. Kraska, E. Upfal, E. Zgraggen, and Z. Zhao. Toward sustainable insights, or why polygamy is bad for you. In *CIDR 2017*.

[7] Y. Katsis et al. Delphi: Data e-platform for personalized population health. In *IEEE Healthcom 2013*.

[8] M. J. Khoury and J. P. Ioannidis. Big data meets public health: Human well-being could benefit from large-scale data if large-scale noise is minimized. *Science*, 346(6213):1054, 2014.

[9] J. A. Morris and M. J. Gardner. Statistics in medicine: Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British medical journal (Clinical research ed.)*, 296(6632):1313, 1988.

[10] M. Santillana et al. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput Biol*, 11(10):e1004513, 2015.

[11] K. Sedig and O. Ola. The challenge of big data in public health: an opportunity for visual analytics. *Online journal of public health informatics*, 5(3), 2014.

[12] T. Sellam and M. Kersten. Cluster-driven navigation of the query space. *IEEE Transactions on Knowledge and Data Engineering*, 28(5):1118–1131, 2016.

[13] T. Siddiqui, A. Kim, J. Lee, K. Karahalios, and A. Parameswaran. Effortless data exploration with zenvisage: an expressive and interactive visual analytics system. *Proceedings of the VLDB Endowment*, 2016.

[14] M. Vartak, S. Madden, A. Parameswaran, and N. Polyzotis. Seedb: supporting visual analytics with data-driven recommendations. VLDB, 2015.

[15] T. Yorifuji et al. Residential proximity to major roads and placenta/ birth weight ratio. *Science of the Total Environment*, 414, 2012.

[16] Z. Zhao, L. De Stefani, et al. Controlling false discoveries during interactive data exploration. *arXiv preprint arXiv:1612.01040*, 2016.