Award Number: W81XWH-13-1-0028

TITLE:
Common Ground: An Interactive Visual Exploration and Discovery for Complex Health Data

PRINCIPAL INVESTIGATOR: Yarden Livnat

CONTRACTING ORGANIZATION:
University of Utah
Salt Lake City, UT 84112

REPORT DATE: December 2015

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:

Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE December 2015	**2. REPORT TYPE** Final	**3. DATES COVERED** 5 March 2013 – 4 Sep 2015

4. TITLE AND SUBTITLE **Common Ground: An Interactive Visual Exploration and Discovery for Complex Health Data**	**5a. CONTRACT NUMBER** W81XWH-13-1-0028
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Yarden Livnat, Per Gesteland, Adi Gundlapalli	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
E-Mail: yarden@sci.utah.edu, per.gesteland@hsc.utah.edu, Adi.Gundlapalli@hsc.utah.edu	**5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) UNIVERSITY OF UTAH, THE 201 S PRESIDENT CIRCLE RM 408 SALT LAKE CITY UT 84112-9023	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The overarching objective of this work is to develop a novel, user-centric visual paradigm aimed at enhancing situational awareness by providing a clear, concise and effective visualization of large, complex and heterogeneous population health data. The project represents collaboration at the University of Utah between the Scientific Computing and Imaging Institute, the Department of Pediatrics, the Department of Medicine and the Department of Biomedical Informatics.

To inform the design of our system, we conducted contextual interviews with experts in the domain of infectious disease epidemiology and population health. We created a database containing over 300,000 encounters (2007-2008) using data from Intermountain Healthcare and developed an ontology for the medical findings in this dataset. We successfully developed a novel software system, named CommonGround, which extends the concept of a disease weather map we developed in a prior project. The system enables users to visually explore and analyze a set of medical findings associated with encounters over a selected period of time. We released CommonGround as a free open source under the most permissive MIT license.

15. SUBJECT TERMS

Visualization, Visual Analytics, Ontology, Situational Awareness, Population Health Data

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	26	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

INTRODUCTION

The overarching objective of this work is to develop a novel, user-centric visual paradigm aimed at enhancing situational awareness by providing a clear, concise and effective visualization of large, complex and heterogeneous population health data. Our aim is to create a flexible and scalable population health visualization that depicts and distills the vast amount of data available from electronic health records using concise meta-data tags. Our goal is to further mature and evaluate an award winning prototype system we developed under the auspices of prior TATRC funding [1][2][3] We hypothesize that a well-designed visualization interface that is tailored to the user's cognitive tasks, supports and promotes the discourse between users and their data and embodies domain knowledge will empower users to actively explore, enhance their ability to comprehend and analyze, and improve overall situational awareness. The project represents collaboration at the University of Utah between the Scientific Computing and Imaging Institute, the Department of Pediatrics, the Department of Medicine and the Department of Biomedical Informatics.

KEYWORDS:

Visual analytics, Situational awareness, Infectious disease outbreaks, Population health

OVERALL PROJECT SUMMARY

The major objectives of the projects were to,

- Extend the concept of a disease weather map to encompass population health data more broadly
- Implement a novel visual analytic system (CommonGround) for use with population health data
- Evaluate usability and impact of the system

Results

We successfully developed a novel software system for enhancing situational awareness that effectively supports visualization of large, complex and heterogeneous population health data sets. The work extends the concept of a disease weather map we developed in a prior project and consists of a completely new code base. To inform the redesign of our original prototype, we conducted contextual interviews with experts in the domain of infectious disease epidemiology and population health. We obtained a large dataset through our collaboration with Intermountain Healthcare containing over 300,000 encounters collected over two years and including a large corpus of clinical concepts abstracted from free-text emergency department dictations, laboratory-confirmed infectious disease cases and the output of a probabilistic Bayesian Influenza case detector. We analyzed and processed the data, extracting 86 unique medical concepts and developing an ontology to represent their semantic relationships. The ontology classifies the medical finding as signs, symptoms, syndromes or pathogen. The findings are also categorized based on main body system such as constitutional, respiratory, gastrointestinal. We created a relational database to store the encounters data and ontology. We developed a new, web-based visual analytics open source system, named CommonGround. The system enables users to visually explore and analyze a set of medical findings associated with encounters over a selected period of time. The visualization is based on a novel interactive tag cloud that provides a global overview of the prevalence of medical findings and facilitates exploration of potential association or correlation between them.

Progress

Elucidate design objectives

To elucidated design objectives and inform the design of the visual interface, we use methods of cognitive task analysis including structured observations and contextual interviews with experts in the domain of infectious disease epidemiology and population health. In addition to the pilot interviews, we conducted interviews with

- Two academic epidemiologists with advanced degrees and extensive experience in population health analysis
- A senior analyst from the Utah State Health Department's Bureau of Epidemiology and Communicable Diseases and the Salt Lake County Health Department
- Two clinician scientists with training in Internal Medicine, Pediatrics, Infectious Diseases and Biomedical Informatics.

The interviews addressed the questions of:

- Discovering cognitive tasks and information seeking intent that can be addressed by CommonGround.
- Defining scope of development, data requirements and ontology
- Informing application tasks and data relationships
- Help define scope of the usability studies.

The interviews also addressed the following technical objectives of the grant:

- Design: Elucidate design objectives. What are the additional needs that are unique to population health, of practicing health professionals, including providers, administrators and planners?
- Data: Develop a scalable approach to deal with the sheer size and complexity of the data through the use of concise and controlled meta-data representation

Data

Both the design and the software development phases in this project were reliant on having access to large and diverse health care data. In the proposal for this work we identified two sources for such data. The first source was the Early Stage Platform (ESP) for Medical Training and Health Information Sciences research and development that was developed by the Advanced Information Technology Group (AITG) in the Telemedicine and Advanced Technology Research Center (TATRC). The second source for data that we identified in the proposal was Intermountain Healthcare, a large, electronically integrated health care delivery system.

ESP Dataset:

The ESP data, described in the RFA for this grant, had great promise as it represented data that is closely aligned with the TATRC mission. The data was to be based on simulated population and thus it would have removed security and privacy concerns. We were unable to get access to the ESP data and we were informed by TATRC that the task of developing the ESP dataset would take much longer than anticipated. Following our first yearly report, where we reported the issue with getting ESP data, we were referred to Ollie B. Gray a research program manager at HITG. We had a conference call on April 28th 2014 and we received small sample dataset containing information for 100 patients on the same day. We spend several weeks processing and analyzing the data. We recruited an undergrad student to write Python scripts to process the data and store it in an RDF repository. After further analysis of the data we came to the conclusion that most of the information is not applicable to our work. Some aspects of the data could in theory be used but after further analysis we concluded that the data resolution and specificity couldn't provide sufficient correlation and detailed insights we need in for this project**.** Demonstrating information visualization capabilities in the domain of communicable disease epidemiology necessitates real world data that contains the dynamic temporospatial features of infectious disease outbreaks. These phenomena are difficult to simulate with artificial data sets. Thus we turned to a collaboration through which we could get real world data.

Intermountain Healthcare data:

The advantage of the Intermountain Healthcare data is that it is based on healthcare operations data from a healthcare system that services a large segment of the population at the state of Utah. Gaining access to such data requires establishing research relationships, negotiating data use agreements, obtaining IRB approval and maintaining compliance with HIPPA regulations. The need for the date and zip code location data for each case patient for this project translates to the need for a limited data set (as opposed to a de-identified data set). We received a final IRB approval from TATRC only on September 10th 2013, more than six months after the start of the project.

During the second half of 2014, Intermountain Healthcare and Dr. Per Gesteland from our team have built a new database containing data for a set of 1,363,464 emergency department visits to six emergency departments in Salt Lake County spanning a seven-year period from 2007 to 2014. The data include basic demographic information (age, postal code, gender), clinical notes (Emergency department physician dictations), microbiological testing results (including testing for respiratory and enteric pathogens that commonly cause outbreaks) and diagnostic information (i.e., ICD-9 codes).

In November 2014, we received a subset of the data containing 1000 records (100 with confirmed influenza and 900 controls) that was used for free-text notes to natural language processing and probabilistic case detection (a Bayesian model for detecting cases of influenza). This dataset is different from traditional data in that it contains both positive and *negative* findings (e.g., "patient denies cough") as well as probabilities based on state-of-the-art Bayesian classifiers. However, the data did not include geographic information or temporal information.

In July 2015, we received the final limited dataset from Intermountain Healthcare. The new dataset included information on 332,284 encounters over two years (2007 and 2008). The information for each encounter consisted of date, age, zip code and a diagnostic information. The diagnostic information per encounter consisted of an abstracted list of clinical findings including signs, symptoms, clinical syndromes and pathogens that were marked as either present, absent or unknown based on natural language parsing of the associated clinical notes. The dataset did not include the actual free text of clinical notes themselves. The dataset also contained a unique kind of data in the form of an estimated probability that the subject of the clinical note had Influenza. The data does not include any identifiers to link encounters with a particular person or indicate if multiple encounters were associated with the same person.

Ontology

Ontological content (terminologies) includes several domains related with patient population health: patient demographic information, geographical information, and clinical information (i.e., diseases, signs, symptoms, infectious agents, and treatments). Instead of a completely new development, we started with an exploration of existing ontological resources, and found good content from these different sources:

- Demographic information was extracted from the Demo-app-ontology developed by William R. Hogan (available at code.google.com/p/demo-app-ontology/). This very detailed and complex ontology was then filtered to focus on the demographic information we would need, and loaded in the Sesame triple store. This filtered ontology includes 35 concepts (i.e., classes) with 18 instances (i.e., named individuals for races and ethnicities) and 9 different relation types (i.e., object properties).
- Geographical information was obtained from the GeoNames geographical database. We extracted detailed information at the state, county, city, and subsidy level for the whole U.S. After converting some content for triple store storage preparation (e.g., converting identifiers from hyperlinks to numerical identifiers), we stored the complete ontology in our Sesame triple store. To allow for efficient querying and use of this resource, we also created a subset focused on the state of Utah. This subset includes 2559 concepts (i.e., classes).

- Clinical information is based on the SNOMED-CT standard terminology. It includes various categories of content (called ‘axis’) such as Body structure, *Clinical finding*, Environment, Event, Observable entity (e.g., age, vital signs, history), *Organism*, Pharmaceutical, Physical object (e.g., furniture, wound dressings), Procedure, Qualifier value, Record artifact, Situation with explicit context (e.g., family history), Social context, Specimen, Staging and scales, and Substance. For this project, we focused on the Clinical finding category for disease and syndromes information, and diagnostic tests. We used the Observable entity category for sign and symptom information. The Organism category provided us with infectious agents information. Finally, the Substance category was used for treatment medications information. The complete terminology was obtained from the National Library of Medicine, and then converted from its original format (called RF 2) to OWL/RDF using an automated script, before loading it into our triple store. Once available in the triple store, we created, analyzed, and visualized subsets of the SNOMED-CT terminology. The first subset included all Glucose metabolism disorders (“Glc Metab Disorder” concept ‘children’). This subset includes 141 concepts with 2 different relation types. The second subset focused on infectious agents from the Procaryote group (e.g., bacteria) and included all relation types. It includes 43927 concepts with 62 different relation types.

Software prototype

The work completed during this project leveraged our work in a previous grant from TATRC in which we developed the EpiCanvas tool. The EpiCanvas tool was developed as a desktop application using Adobe Flex framework (www.adobe.com/products/flex). Our aim in this project was to develop a web-based application that could be accessed from any modern web browser. To this end, we employed a client-server architecture in which the server is responsible to all the communications with the data repository. The advantage of this approach is the decoupling between the client and the data repository, which means that the client does not depend on a specific type of data repository implementation. It also reduces security concerns as the data repository is kept behind a firewall (i.e., the data is stored locally) and is accessible only by the dedicated server. The new architecture comprises three layers: the data layer, middleware and a presentation layer.

In the proposal, we anticipated receiving data from the ESP project or from Intermountain Healthcare within 3 to 6 months after the start of the project. Both the ESP project and the Intermountain Healthcare internal projects were far behind schedule. We received a small sample data from Intermountain Healthcare only in November 2014 and a final dataset in July 2015. The final dataset was very different in structure and content from the first sample data. The staggered availability and internal inconsistencies of these data sets led us to develop three different software prototypes.

First software prototype

We developed the first software prototype based on the assumption that the data we will receive will be similar of the data we received from Intermountain Healthcare in the past. We evaluated various options for storing data such that we will be able to issue queries relating to both the health data and to the knowledgebase at the same time. In particular, we looked at RDF repositories such as the open source Sesame (rdf4j.org), traditional relational databases and NoSQL type databases (nosql-database.org). We’ve used Sesame in previous projects to store ontologies and RDF based data but we concluded that Sesame might not scale well enough and that it doesn’t offer the flexibility we need. We also looked at various types of NoSQL databases: key-value stores, column stores (e.g. Cassandra cassandra.apache.org), document databases such as MongoDB (mongodb.com) and graph databases such as Neo4J (neo4j.org).

We chose to implement our data repository using a graph database and specifically on the Neo4J implementation. A graph database is a form of a NoSQL database, which have recently gained popularity in big data and real-time web applications due to their simplicity, scalability and support for finer control. Since we did not have data from Intermountain Healthcare at that time, we used data we had from phase I of the project. We designed a graph-based representation for the data, converted the data, meta-data and the knowledgebase to this format and finally deployed the new graph based repository on our server. We also created two additional graph databases for the ESP data and for the initial data from Intermountain Healthcare. Developed another graph database for encoding and storing initial data from Intermountain Health. The new data incorporate ICD-9 codes and we are working on pruning the data and converting ICD-9 codes to short text labels appropriate for our display.

We developed the first software prototype using AngularJS, an open source framework by Google (angularjs.org), which has gained popularity for web development over the last few years. However, as AngularJS gained more popularity, concerns have been voiced about the platform's steep learning curve, performance limitations and architectural complexity. These challenges proved to be major obstacles for us as well, in part because there was no prior experience with it at the University of Uth research environment. During development, we found out that most of the open source software components for AngularJS were of low quality, poorly maintained and very limited in their capabilities. Despite these challenges we were able to successfully develop an AngularJS based visual analytic interface, albeit the development cycle was much longer than anticipated.

Second software prototype

In November 2014, we received a small sample dataset from Intermountain Healthcare, which was different in several aspects from the data we had previously worked on. The sample dataset had no geographic information and no dates though it had new probabilistic indications. As such it did not fit the assumptions of the first software prototype and we had to develop a second prototype. At that stage, Intermountain Healthcare had not finalized at that stage the final format and scope of the data they intended to provide us. For this reason, we opted to store the sample data in a standard relational database (MySQL) and forgo the use of the Neo4J data repository.

In late 2014 Google announced that version 2 of AngularJS would not be back compatible with AngularJS 1.x and will represent a major architecture change. Due to the issues with AngularJS and the need to develop a new user interface we developed the second software prototype from scratch using only the D3 graphics library (http://d3js.org). As part of the new code we developed visual encoding to show both positive and negative association between tags and cases. We also enhanced the filtering capabilities to enable users to also filter data by explicitly excluding tags. This allow users to express a filter such as "all encounters that are associated with `Chill` but not `Sore Throat` or `Nausea`".

Final software prototype

In July 2015 we received the final dataset from Intermountain Healthcare. Since this dataset was different from what we had initially expected and from the first sample dataset we received from Intermountain Healthcare it necessitates the development of a third software prototype. In this case, we were able to leverage the code from the second prototype and incorporate it into the final code. The final software incorporates several new features including geographic map depicting spatial distribution of encounters, historical patterns of various pathogen infections and a tags ontology. We developed new software to incorporate the new probabilistic case detection data and new visual encoding of these probabilities. In the previous two prototypes, users could filter data only via the tag cloud interfaces by selecting various visible tags. For this prototype we designed a new filtering system that allows users to filter data by several different methods including selecting specific zip codes on the map, selecting various findings or topics from the tag ontology and selecting specified probabilistic thresholds using results from the Influenza case detector. We developed a new correlation measure that facilitates time-based correlation between meta-data tags that are mutually exclusive (such as two cities or two age groups).

Evaluation

The third aim of the project was to conduct user studies to evaluate usability and impact of the system. The expectation was that we will receive appropriate data within the first 3-6 months and be able to conduct and evaluation in the last two months of the one-year project. These expectations did not materialize. Intermountain Healthcare provided us with data almost two years after the expected timeframe and only two months before the extended end of the project (after receiving two no-cost extensions). The unexpected type of data also necessitates the development of a new software prototype. These delays and challenges have left no time to conduct extensive evaluation of the software.

Accomplishments and discussion

Extending the concepts of a disease weather map

To inform the redesign of the original CG prototype we conducted contextual interviews with experts in the domain of infectious disease epidemiology and population health. A report detailing the findings of these interviews including the concept categories and evidence to support of their creation is provided in Appendix A. As part of this project we collaborated with Intermountain Healthcare, which is developing a case detection system predicated on free-text emergency department clinical notes. The Intermountain Healthcare effort is done in parallel through another funding source. In the next section, we synthesize the major conceptual extensions elucidated from this this analysis with our extant ideas for extending our work summarize the needed functional enhancements.

Incorporating Probabilistic Case Detection and Uncertainty

Effective surveillance and situational awareness in the domain of public health and infectious disease epidemiology is premised on the ability to detect cases of the disease(s) or condition of interest. The paradigm of case detection, in this context, needs to account for the uncertainty inherent to the data available to a case detector or detection method. In clinical settings, the specificity of data available for detection can range from non-specific signs and symptoms exhibited during illness to highly specific laboratory confirmation of a pathogen (e.g., molecular detection of a pathogen and quantitative assessment of a host serological response). Case detection, is thus, a probabilistic endeavor whereby a human (e.g., a public health official reviewing medical records) categorizes potential cases – *clinically compatible case, confirmed case, epidemiologically linked case, laboratory-confirmed case, probable case, suspected case* – after reviewing the available data and applying an accepted case definition [5].

The state of the art in infectious disease case detection is increasingly taking advantage of the rich data available in electronic health records, including the presence or absence of clinical findings historically locked away in free text clinical notes. Utilizing several technologies, including natural language parsers and machine learning methods (e.g., Bayesian networks, support vector machines, Gaussian mixture models), probabilistic case detectors have been developed for influenza [6] and are being developed for other diseases. In contrast to human-based categorical approaches described above, these systems automatically compute prior probabilities for the disease of interest and output a continuous variable (between 0 and 1) that can be employed to explore case detection across a spectrum of certainty. For example, a user might want to only see cases for which there was a high probability of disease (e.g. > 90%). Further, as additional case detectors are developed (i.e., for various pathogens/conditions of interest), surveillance visualization systems will need to provide intuitive ways for the user to interact with the data through these novel lenses.

We incorporated the ability to visualize prior probabilities for Influenza and by taking advantage of a case detection system predicated on free-text emergency department clinical notes. This system was being developed, in parallel, through another funding source at the same time as our work. Figure 1 depicts two screen shots from the CommonGround interface. On the left, the user has selected `cough`, `sore throat`, `fever` and `ILI`. On the right, the user has selected `runny nose` and `cough`. The distribution of encounters vs. probability (0.2 to 1 on the X-axis) of having Influenza is dramatically different in each situation.

The system is extensible and can support data exploration through the lenses of multiple case detectors using the same substrate data. As effective case detectors are developed for specific pathogens (e.g., certain viruses, atypical bacteria) and clinical syndromes (e.g., Zika-like, enterovirus-like illness), they can be readily integrated into the CommonGround interface.

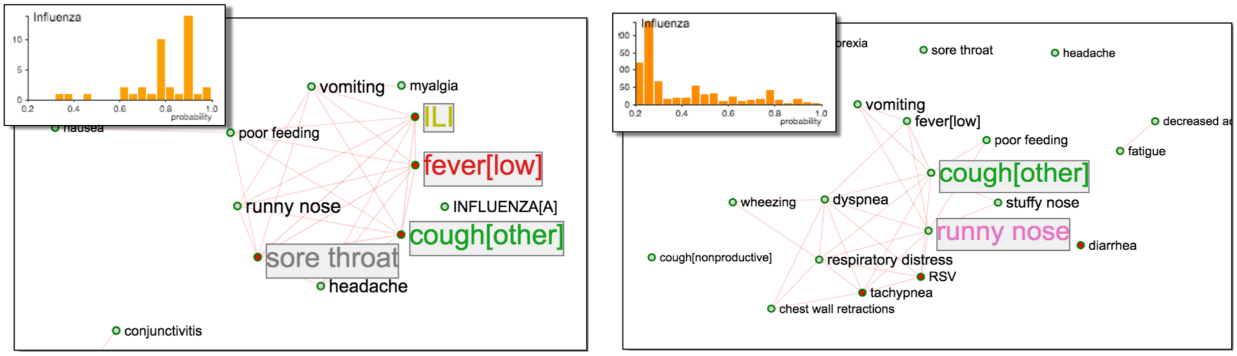
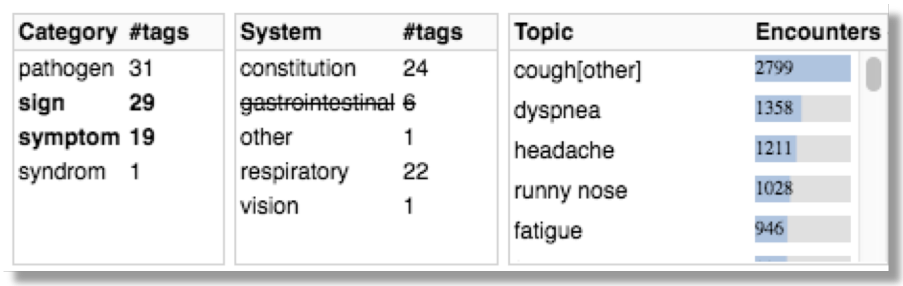


Figure 1 Distribution of encounters vs. probability(x-axis) of having Influenza for encounters with cough, fever, sore throat and ILI (left) compared to encounters with cough and runny nose.

Representation of Domain Knowledge

The ability to incorporate domain knowledge into information visualization systems represents an important contribution from the fields of computer science and artificial intelligence to biomedical research and innovation. From an informatics perspective, integrated medical/biological ontologies and new semantic-based models for managing information provide both new challenges for research and opportunities for translation into the practice of epidemiology. We developed an ontology to provide semantics to the tags based on the final dataset we received from Intermountain Healthcare. The ontology contained 86 concepts (tags) representing symptoms, syndromes, signs and pathogens. The ontology also describes five different systems: constitution, respiratory, gastrointestinal, vision and other. The ontology is described in Appendix C. In addition to incorporating the ontological relationships into the queries populating the CommonGround interface, we implemented a set of ontologically driven filters that make it possible for the user to explore the by constraining the view to different concepts at different levels semantic representation (i.e., by signs, symptoms, syndromes, specific pathogens). Figure 2 shows a screen shot from the CommonGround interface depicting the various classifications available in the data. From this top-level control, the user can select or exclude various categories and specific tags. For example, in Figure 2 the user is focusing only on signs and symptoms that are not associated with gastrointestinal.



Category	#tags
pathogen	31
sign	**29**
symptom	**19**
syndrom	1

System	#tags
constitution	24
~~gastrointestinal~~	~~6~~
other	1
respiratory	22
vision	1

Topic	Encounters
cough[other]	2799
dyspnea	1358
headache	1211
runny nose	1028
fatigue	946

Figure 2 Example of an ontology driven filtering in which the user created the filter "signs and symptoms but not gastrointestinal."

We have been able to extend our concept of an infectious disease weather map in several important ways. First, we were able to integrate additional data types (dimensions) into the systems. Multidimensional data is a critical feature to consider when developing data visualization tools for most domains but especially in the, somewhat messy, biomedical domain. Second, we were able to develop a pilot ontology to demonstrate the value of this rapidly emerging technology for information visualization and knowledge representation. This experience has provided valuable insights into how we could take advantage of production ontological services to help organize data and integrate domain knowledge for the CG approach.

Data

We created a relational database with longitudinal data consisting of 332,284 encounters over two years (2007 and 2008). Each encounter consists of date, age, zip code and diagnostic information about signs, symptoms, syndromes and pathogens. The database also contains unique information from a novel Influenza case-detector in the form of probability information per encounter. Appendix B describes the database schema.

Software prototype

We developed a flexible and scalable population health visualization web-based software prototype, named CommonGround, that depicts and distills the vast amount of data available from electronic health records using concise meta-data tags. The visualization provides a dynamic display of the active tags (emphasizing tags based on their current relative importance) and emerging temporal correlations between the tags.

System backend

The system backend consists of a web server based on the open source Node.js (nodejs.org) and Express.js, an open source web framework for Node.js. The web server provides an access point for users and a gateway for the client to query the database without direct access to it. The CommonGround client sends generic queries to the server, which in turn connects to the database and issue the appropriate SQL query. The server also hosts public demographic information in the form of population size per zip code.

Tags association metrics

We created several measures to define the strength of connection between each pair of tags. The measures assign a value in the range of [0, 1], where 0 means no relation and 1 means strong relation.

Symmetric association: An association based on the number of shared encounters between the the two tags. If tag *T1* is associated with a set *E1* of of encounters and *T2* is associated with set *E2* than we define the association A_1 between *T1* and *T2* as

$$A1(T1,T2) = \frac{|E1 \cap E2|}{|E1 \cup E2|}$$

where |X| represents the size of set X. The association A_1 is thus symmetric and we represent it as a line between the two tags.

Directed association (suggest): If the size of *E1* is much larger than the size of *E2* (or vice versa) than the association A_1 between them is be definition very small even if most of the encounters in *E2* are part of *E1*. For example in the extreme case where all the encounter associated with *T2* are also associated with *T1* than $E2 \subset E1$ and $A1(T1,T2) = |E2|/|E1| << 1$. In such cases, measure A_1 doesn't provide sufficient insights. To address this, we defined a second measure of association as

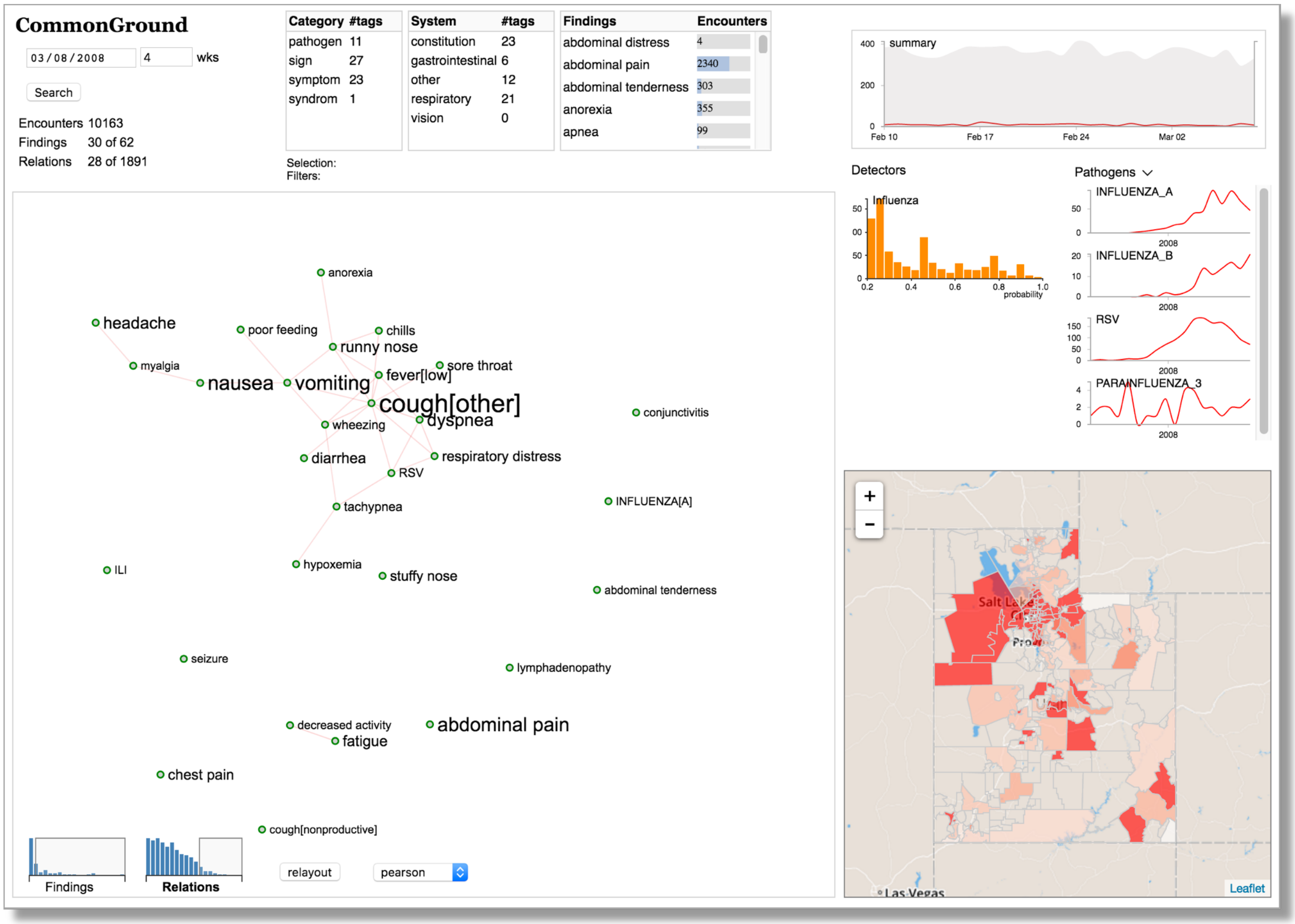
$$A2(T1,T2) = \frac{|E1 \cap E2|}{\min(|E1|,|E2|)}$$

Measure A_2 describes the relative importance of the shared encounters relative to the smaller set and can be thought of as one tag (smaller set) *suggest* a second tag (larger set). When using measure A_2, we depict the relationship between two tags as directional arrow from the *T2* to *T1* (assuming E1 » E2).

Correlation: A correlation measure based of Pearson correlation. Instead of using the number of shared encounters between each pairs of tags, we consider the time series of the number of encounters per day for each tag and compute the Pearson correlation between these time series. One advantage of this measure is that it can expose correlations between mutually excluding tags, such as between geographical tags (different cities or zip codes) and between age groups (e.g. children, adults) that by definition can't have any shared cases.

Coordinated multiple views

An important aspect of a visual analytics system is the ability to view different aspects of the data side by side and coordinate between these views, also known as Coordinated Multiple Views [7]. Figure 3 shows a screenshot of the CommonGround visual analytics client.



CommonGround

03/08/2008 4 wks

Search

Encounters 10163
Findings 30 of 62
Relations 28 of 1891

Category	#tags
pathogen	11
sign	27
symptom	23
syndrom	1

System	#tags
constitution	23
gastrointestinal	6
other	12
respiratory	21
vision	0

Findings	Encounters
abdominal distress	4
abdominal pain	2340
abdominal tenderness	303
anorexia	355
apnea	99

Selection:
Filters:

Figure 3 The CommonGround visual analytics display showing medical findings over four weeks. The findings are scaled based on their relative importance and arranged based on the strength of their pairwise Pearson correlation (strong correlations are shown as red edges). The geographic map shows rate of encounters per 1000 people per zip code.

The display comprises of a collection of views organized in four main areas:

- ***Query*** area, top left, enable users to specify the period of time of interest (start date and length in weeks). The CommonGround system then query the database and identifies all the encounters that occurred during this period of time and the tags associated with each encounter.
- ***Tag Cloud*** area, bottom left, depicts the tags associated with the current list of encounters and temporal relationship between them. The *Tag Cloud* display is the main component of the CommonGround visual analytics display. The size of each tag represents the number of reported cases associated with that tag. The location of the tag on the screen depends on the temporal association between the tags.
- ***Tags Information*** area, top middle, provides general information about the extracted tags based on their semantic. The first two tables depict the number of tags for each tag category and system respectively. The third table list all the extracted tags and the number of extracted encounters associated with them. The Encounter column also depicts the histogram of the number of encounters per tag (blue areas). The user can sort each table based on either column and in ascending or descending order.

- ***Encounters Information*** area, right, depicts various information on the selected encounters. The top chart shows a summary of the number of cases per day. A geographic map, based on the Leaflet (http://leafletjs.com) open software JavaScript library, depicts a heat map of the number of encounters per zip code. Historical information about various pathogens prevalence is shown in a collection of epicurves on the right. The epicurves always depict the last 6 months. Using a pull-down menu (Pathogens) enables users to select which pathogen epicurves to show.

The system maintains two lists of encounters. The first list consists of all the encounters that fit the user query (date and period) and that were fetched from the database along with their corresponding tags. The second list consist of *filtered* (or *active*) encounters that correspond to the encounters that passed the user's filters if any. By definition, the default filter (the case where no other filters were provided) includes all the encounters in the first list. The various views depict different aspects of the *active* encounters and the tags associated with them.

Tag Cloud view

The tag cloud view depicts the collection of tags associated with the active encounters list. These active tags form a graph where each node in the graph represents a tag. Edges in the graph represent a temporal relation between the corresponding tags. The user can select which measure to use such as association or correlation. Given such a graph we use a force-directed graph layout to arrange the tags on the screen. The size of each tag in the tag cloud depends on its relative importance, which we calculate based on the number of encounters that are associated with that tag. We define the importance $I_{i,}$ of tag T_i as the number of encounters associated with tag T_i

$$Ii = \sum_j \begin{cases} 1 & if\ ej\ has\ tag\ Ti \\ 0 & otherwise \end{cases}$$

where e_j represents encounter j and the sum is computed over all the encounters. We then define the relative importance RI of tag T_i as

$$RI(Ti) = \frac{Ii}{\max_k(Ik)}$$

where the max is computed over all tags.

To help reduce screen-clutter and enable analysis of the graph, the tag cloud view incorporates two histograms at the bottom left of the view. One histogram depicts the number of tags (y-axis) vs. tag relative importance (x-axis). A second histogram depicts the number of graph edges (relations between pairs of tags) vs. edge strength (x-axis). Both histograms use 20 bins over the x-axis (range 0-1). The user can focus on a subset of the tags by brushing over the first histogram to specify a range of values. Only tags with relative importance within this range are then shown. Similarly, the user can focus only on edges (relationship) within a specific

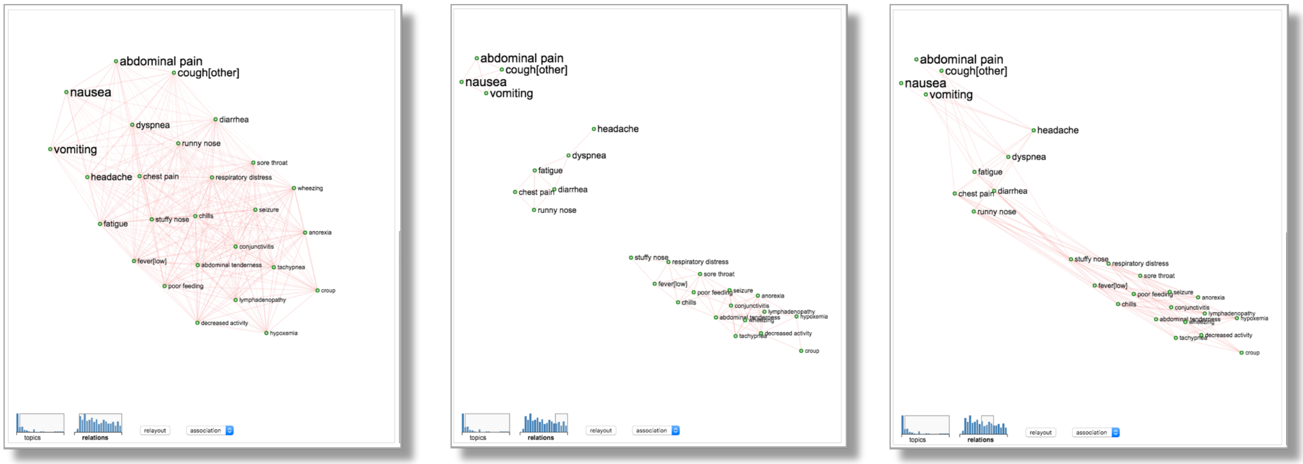


Figure 4 Tags layout using all relations between the tags (left). Using only strong relations creates three distinct clusters (middle). Right: examining inter-cluster relationships by utilizing the middle layout but focusing on middle range relations can show inter cluster dependency.

range by brushing over the second histogram. Note that brushing over the histograms only determines which tags and edges are shown. It does not filter the data and the information in the other views doesn't change.

Further analysis of the tag cloud can be achieved by brushing over the histogram and the requesting the system to reapply the force-directed graph using the 're-layout' button. Although brushing does not affect the underlying graph, we apply the force-directed layout only to the visible nodes and edges. In many situations there are many weak relations between the tags because they may share only a few encounters. Although these relations are weak they do constrain the layout arrangement of the tags in the display. Filtering out weak relations enables the layout algorithm to better cluster strongly related tags as show in the middle image in Figure 4**.** After the tags are cluster based on strong relations, the user can explore weaker relations by selecting a larger or a different range of the histogram while maintaining the same layout (and thus clustering) arrangement.

Case detectors

The CommonGround prototype supports multiple cases detectors, although our data included from one Influenza case detector. The list of available cases detectors in the databased is shown on the left side in the *Encounter Information* area. A case detector assigns to each encounter the probability the encounter fits the case detector case definition. In our case, the Influenza case detector was trained on laboratory confirmed cases of Influenza. Each case detector chart depicts a histogram of the number of encounters (y-axis) vs. their probability (x-axis). Users can use case detector histograms to modify the tag cloud display and filter encounters.

To incorporate case-detector probabilities we modify the importance function to be the weighted sum of the probabilities

$$I_i = \sum_j \begin{cases} p(e_j) & if\ e_j\ has\ tag\ T_i \\ 0 & otherwise \end{cases}$$

where $p(e_j)$ is the probability the case detector assigns to encounter j.

Selecting a particular case-detector effectively changes the relative importance of each tag. For example, using the Influenza case-detector will likely increase the importance of tags such cough and headaches but reduce the importance of abdominal pain and diarrhea, as shown in Figure 5. Note that the relative importance of the tags

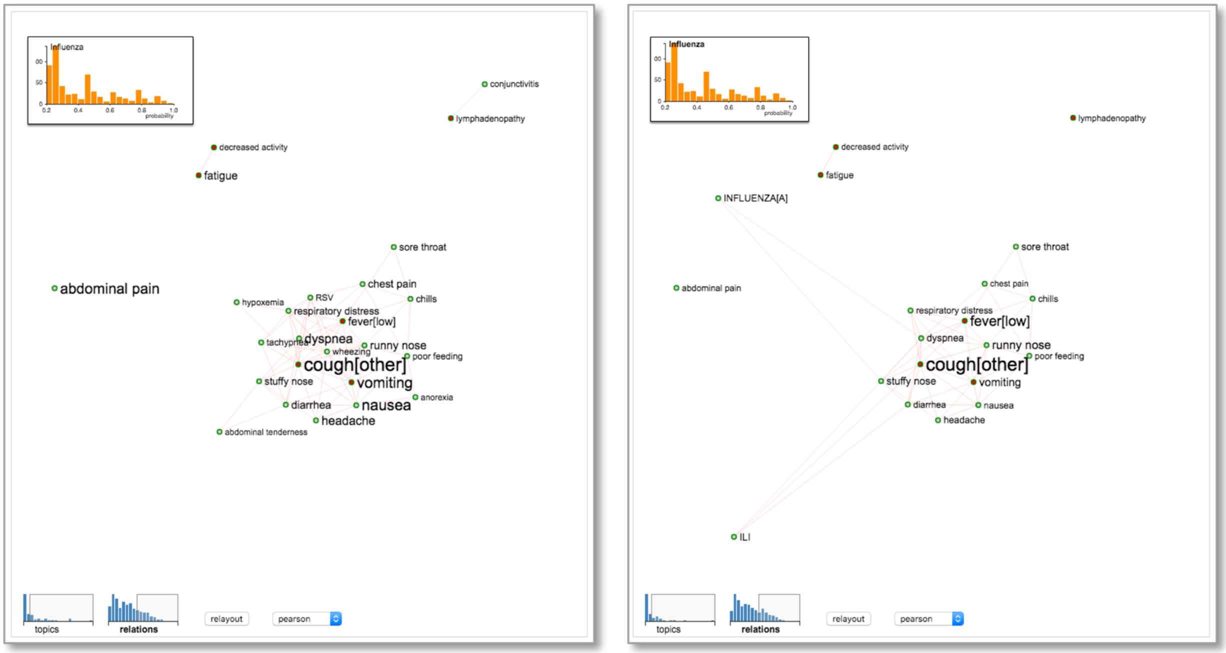


Figure 5 Comparison of the tag cloud using the default importance function (left) vs. modifying the importance based on the probability of an Influenza case detector (right). The relative importance of fatigue, vomiting, nausea and abdominal pain decreases while fever and runny nose increases. Note that conjunctivitis and hypoxemia disappeared all together as their relative importance fell below the user threshold, while tags such as ILI and INFLUENZA[A] emerged.

depends on the probabilities assigned by the case detector to each encounter and not on a pre-determine semantic of the tags. In this way, the display reflects the instantaneous important of the tags. If unexpectedly vomiting is prominent with people who are likely to have the flu than the relative importance of the vomiting tag will increase and the size of the tag will be much larger than usual.

The case-detector data is also used to filter the data by brushing over the histogram to select a range of probabilities assigned by the detector and focus only on encounters with probabilities in that range. Detector based filtering can facilitate identifying symptoms and other findings that are likely temporally associated with the case definition. Conversely, focusing on low probabilities can help identify findings that are not likely to be part of a potential outbreak.

Filtering using tag selection and exclusion

In addition to promoting situational awareness, the tag cloud also enables users to drill down into the data through interactive filtering. Selection of a tag (a click on the tag) amounts to selecting only the encounters with that particular tag. Selection of multiple tags is equivalent to a logical AND, i.e. selecting the encounters that have all of the tags. We extended the tag based filtering capabilities to enable users to express exclusion (a logical NOT), which amount to selecting only encounters that do not have any of the excluded tags. Selection and exclusion are performed by clicking on a tag or clicking while pressing the Meta/Alt key. In the second prototype we distinguished between selected vs. excluded tags by encoding selected tags with a gray background and excluded tags with a yellow background. This color encoding was not compatible with other visual encoding we employed in the final prototype. Instead, we encode with a gray background any tag that is currently being used for filtering (selected or excluded). Excluded tags are further marked with a strike through as is shown in Figure 6.

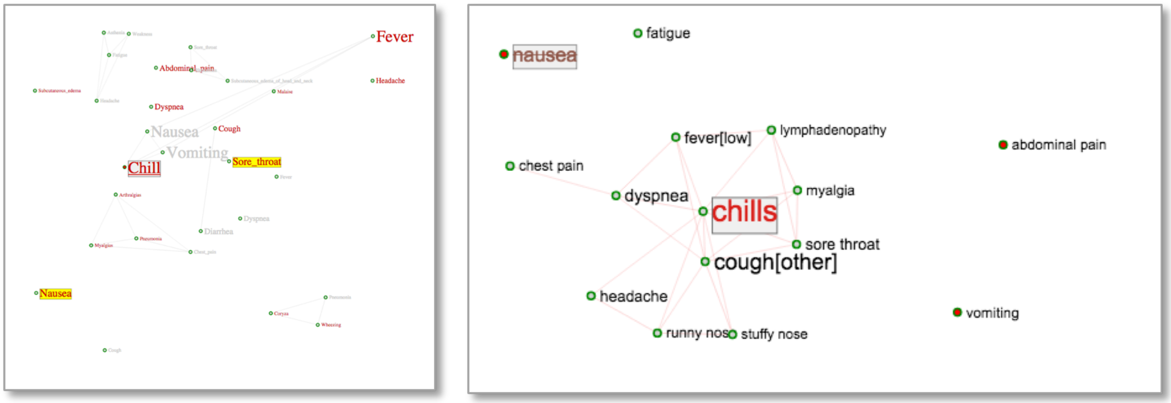


Figure 6 Filtering using tag selection and exclusion. Selected tags are visually encoded using a gray background. Excluded tags were encoded using a yellow background in the second prototype (left) and using a strikethrough font in the final prototype (right).

Confirmed vs. absents findings

In general, a tag represents a positive finding associated with an encounter. According to the open world assumption a *lack* of a finding *doesn't* mean negative finding. In some cases, the clinical notes associated with an encounter may explicitly state that a particular finding was absent, for example no fever. One of unique features of the data we received from Intermountain Healthcare is occasional explicit indication of the absent of a finding. We experimented with visualizing negative findings by visually encoding positive finding in red and negative finding in light gray. The left image in Figure 6 depicts a screenshot of the user interface of the second software prototype in which the user constructed the filter "chill and not sore throat and not nausea". The display shows that many encounters with "chill and not sore throat and not nausea" also have fever but not vomiting. For the situation depicted in Figure 6, most of the encounters that were not associated nausea were also explicitly marked as not having `nausea` (negative finding) but the same is not true for `sore throat` (note the lack of negative `sore throat`). The capability to show explicit absence of a finding can potentially be of value for case ascertainment (i.e., determining whether a case meets or does not meet a case definition) but the lack of consistent assessment of the absence of a specific finding in actual clinical practice (as opposed to an official public health case investigation) reduces the trust public health official places in such information and can cause confusion. In addition, negative findings can be attributed much more often than positive findings, which can diminish the relative importance of the positive findings. Due to these concerned we opted not to include negative finding in the final software prototype.

Visualization relationships between tags

Pairwise association or correlation between tags, are visually encoded as edges between the textual tags. In the case of directional association, we use arrows instead of regular lines to provide users with a clear visual cue that one tag *'suggest'* a second tag, i.e. if an encounter has one tag than it's likely that it will also be associated with the second tag, as shown in Figure 7. We experimented, in the second software prototype, with using curved edges to reduce overlap between edges and labels the tags as shown in Figure 7. Curved edges exhibit a nicer overall look but our experiments reviled that they are harder for users to comprehend. A direct edge emanating from a tag gives a clearer indication of where the target tag might be located on the screen. A curved edge, on the other hand, forces the user's gaze to follow the edge all the way from one side to the other and thus is not well suited for overviews that are essential for situational awareness.

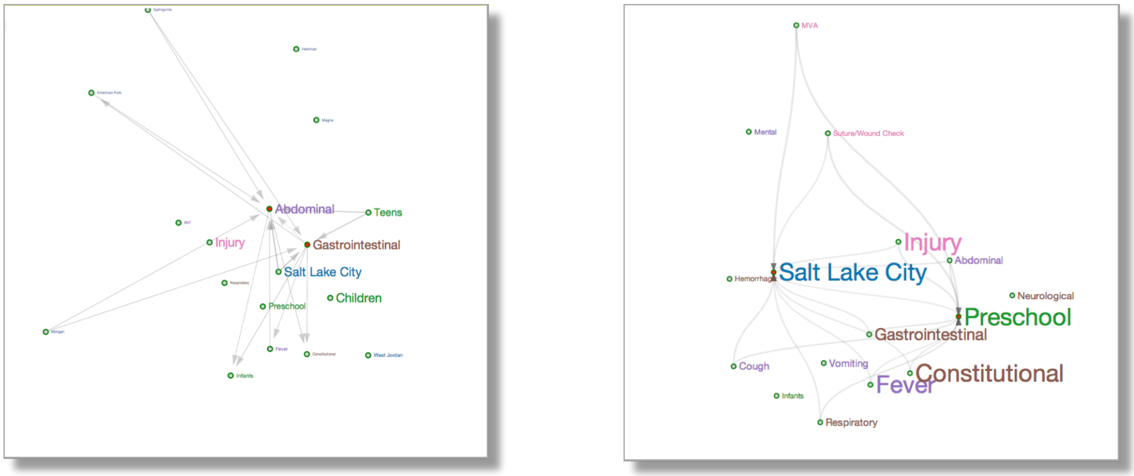


Figure 7 Visualizing directional associations (tag1 suggest tag2) and using curved edges to reduce overlap between edges and tags (right).

Cross Filtering

We developed a novel filtering system based on the concept of a Crossfilter (square.github.io/crossfilter). A crossfilter provides a fast multidimensional filtering for coordinated views system. However, the Crossfilter library supports only a single data table and only with a small number of columns. In contrast, our system must deal with multiple data tables (encounter, tag semantics, case detectors) and with a many-to-many table describing relations between encounters and tags, none of which Crossfilter can address.

We designed a filtering system consisting of a collection of tightly integrated crossfilters and developed special code to handle the encounters to tags many-to-many relations. The filtering system supports interactive data filtering through selection of multiple attribute values in different views. When a filter is modified in one view the system re-filter the encounters list and refreshes all the views to reflect the new active list. An important aspect of the filtering system is its ability to show in each view both the current active list (after applying all the filters from all the views) and an active list based on all the filters *except* the filter associated with that view. For example, the geographic map shows the number of active encounters per zip code and allows the user to filter the data by selecting multiple zip codes. Consider what happens after the user select the first zip code. If the map will reflect the newly filtered data, then only the selected zip code will show any active encounters, thus preventing the user from selecting other relevant zip codes.

Using our system-wide cross filtering capability we enable the user to apply a variety of filters:

- *Geographic map:* filter the active encounters based on one or more zip codes
- *Tag category:* selecting one or more tag categories shows only tags associated with these categories and filters only encounters that have these tags.
- *Tag system:* selecting one or more tag systems shows only tags associated with these categories and filters only encounters that have these tags.
- *Tag Cloud:* select or exclude a set of tags. Filter the active encounters list to include only encounters that has the selected tags but do not have any of the excluded tags.
- *Topic list:* Same as the Tag Cloud.
- *Detector:* When a detector is active the user can select a rage of probabilities. Filters encounters with probabilities in that range.

KEY RESEARCH ACCOMPLISHMENTS

- Created a graph database containing sample dataset from Intermountain Healthcare
- Created an SQLite database containing 332,284 encounters over two years (2007 and 2008) and probability information from an Influenza case-detector
- Developed an ontology for a collection of 86 signs, symptoms, syndromes, and pathogens.
- Developed a web-based prototype using AngularJS and D3 and a web server using Node.js.
- Developed a second software prototype to take advantage of new attributes in the second dataset we received from Intermountain Healthcare.
- Develop a final software prototype to handle new longitudinal data, incorporate infectious disease case detectors, historical lab confirmed pathogens and sematic tags.
- Gave a presentation at Goldman Sachs in Salt Lake City that was broadcasted live to their offices in Austin Texas, NYC, London, Bangalore and a second office in Utah.
- Received an NSF SBIR grant ($750K over 2 years) that is based in part of the exploration display approach we developed under this grant.

CONCLUSION

The overarching objective of this work is to develop a novel, user-centric visual paradigm aimed at enhancing situational awareness by providing a clear, concise and effective visualization of large, complex and heterogeneous population health data. The project represents collaboration at the University of Utah between the Scientific Computing and Imaging Institute, the Department of Pediatrics, the Department of Medicine and the Department of Biomedical Informatics.

To inform the design of our system, we conducted contextual interviews with experts in the domain of infectious disease epidemiology and population health. Although obtaining appropriate data proved to be a challenge and introduced delays, we were nevertheless able to obtain a large dataset from Intermountain Healthcare containing over 300,000 encounters collected over two years. We also developed an ontology for the medical findings in this dataset.

We successfully developed a novel software system, named CommonGround, for enhancing situational awareness by effective visualization of large, complex and heterogeneous population health data. The work extends the concept of a disease weather map we developed in a prior project and consist of a new code base. The system enables users to visually explore and analyze a set of medical findings associated with encounters over a selected period of time. We publically released CommonGround as a free open source under the most permissive MIT license.

The CommonGround system incorporates various measures to compute the relative importance of a medical finding and relationships between them. Additional measures can be developed to provide additional insights such as measures based on importance relative to past history or relative to the category or system of a medical finding. We were unable to conduct an evaluation study of the software due to the delay in obtaining data. Nevertheless, we intend to conduct this study outside of the scope of this grant in the near future. We already applied and received an IRB extension from the University of Utah to continue this research.

Our results and deliverables are based on applying sound informatics techniques and principles to visualize and explore large, complex and heterogeneous population health data will serve as viable models for analysis of big healthcare data. We are working now with Intermountain Healthcare to incorporate a variation of this software into their new GermWatch project. The work has also attracted additional interest from public health officials in Ontario Canada. While the specific domain of interest to this project is bio-surveillance with a focus on respiratory infections, we note that the scientific principles of user-centric design and contextual inquiries are portable to other domains. The work has been incorporated in part into two NSF SBIRs grants and received interest from Pacific Northwest National Lab, Goldman Sachs and several other commercial companies.

PUBLICATIONS, ABSTRACT, AND PRESENTATIONS

PUBLICATIONS:

- Y. Livnat, E. Jurrus, P. Gesteland, A. V. Gundlapalli," The CommonGround Visual Paradigm for Biosurveillance", *Workshop on Signature Discovery, Intelligence and Security Information*, pp. 352-357, June 2013.

PRESENTATIONS:

- Y. Livnat, Visual Exploration for Situational Awareness, *Goldman Sachs*, SLC (live broadcast to London, Bangalore, NYC and Austin Texas), 2015
- Y. Livnat, The CommonGround Visual Paradigm for Biosurveillance, *Pacific Northwest National Lab (PNNL)*, 2013
- * Y. Livnat, The CommonGround Visual Paradigm for Biosurveillance, *Workshop on Signature Discovery, Intelligence and Security Information*, Seattle, 2013

INVENTIONS, PATENTS AND LICENSES

- Intermountain Healthcare is in advance negotiations with the University of Utah to license the CommonGround software to incorporate it into their new GermWatch project.
- Enclavix, Inc. is working on an SBIR project with the University of Utah that incorporate some of the code from CommonGround. As part of this effort, Enclavix has agreed to license the technology from the university at the end of this development effort.

REPORTABLE OUTCOME

Database: Created a database with longitudinal data consisting of 332,284 encounters over two years (2007 and 2008). Each encounter consists of date, age, zip code and a diagnostic information about signs, symptoms, syndromes and pathogens. The database also contains unique information from a novel Influenza case-detector in the form of probability information per encounter. Due to HIPPA restrictions and our data shared agreement with

Intermountain Healthcare, we can not release this data. The public repository of our software prototype does *not* include this data.

Ontology: Developed an ontology that represents semantic information for 86 medical findings. The ontology contains classifications based on *category* (symptoms, syndromes, signs and pathogens related to Influenza Like Illnesses (ILI)), *system* (e.g. respiratory, vision) and additional details information.

Software prototype: Developed an interactive visual analytics web based software prototype. The open source software is freely available on GitHub at https://github.com/yarden-livnat/cg under an MIT license, which is the most permissible license and allow anyone to use or modify the code without restrictions or fee.

OTHER ACHIEVEMENTS

Degrees: The following students received an M.Sc. from the University of Utah and where partially funded by this grant.

- Vineel Yalamarthy received an M.Sc. in 2015
- Mounika Reddy Kodur received an M.Sc. in 2014
- Narong Boonsirisumpun received an M.Sc. in 2014

Funding: The overall approach we developed as part of this grant is general and is based on abstract concepts rather than domain specific. The software prototype we developed is specific to public health data, yet we were able to apply the approach to other domains. We received the following NSF funding based on this work

- NSF, SBIR Jan 2014-June 2014
  Budget: $150K (UofU: $50K)
  PI: Nicole Davis (Enclavix), Yarden Livnat (Co-PI, UofU PI)
  Title: Create a Machine Learning-based system to Educate and Support Entrepreneurs
- NSF SBIR Sept 2014 – Aug 2016
  PI: Brad Davis (Enclavix), Yarden Livnat (UofU PI)
  Budget: $750K/2 years (UofU: $375K/2 years)
  Title: Automated System to Identify and Curate Web-based Resources for Entrepreneurs

REFERENCES

[1] A. V. Gundlapalli, Y. Livnat, and P. H. Gesteland, “Final Report Submitted to U.S. Army Medical Research and Materiel Command: Visual Correlation for the Early Detection of Infectious Disease Outbreaks Award Number W81XWH0710699)”, *University of Utah School of Medicine*, Salt Lake City, UT, 2009.

[2] P. Gesteland, Y. Livnat, N. Galli, M. H. Samore, A. V. Gundlapalli. “The EpiCanvas Infectious Disease Weather map: An Interactive Visual Exploration of Temporal Correlations”. *Journal of the American Medical Informatics Association (JAMIA)*, Vol. 9, pp. 954-959, 2012. [ISDS Award for Outstanding Research Article in Biosurveillance (Scientific Achievement) selected from all journal publications since 2010, ISDS 2012

[3] P. Gesteland, Y. Livnat, N. Galli, M. H. Samore, A. V. Gundlapalli. “The EpiCanvas Infectious Disease Weather map: An Interactive Visual Exploration of Temporal Correlations”. *AMIA 2011 Annual Symposium*, 2011 (Winner of the Homer R. Warner Award)

[4] Y. Livnat, E. Jurrus, P. Gesteland, A. V. Gundlapalli,” The CommonGround Visual Paradigm for Biosurveillance”, *Workshop on Signature Discovery, Intelligence and Security Information*, pp. 352-357, June 2013.

[5] Case Definitions for Infectious Conditions Under Public Health Surveillance, *MMWR Recommendation and Reports*, 46(RR10), 1-55, May 1997 (http://www.cdc.gov/mmwr/preview/mmwrhtml/00047449.htm)

[6] Fuchiang Tsui, Michael Wagner, Gregory Cooper, Jialan Que, Hendrik Harkema, John Dowling, Thomsun Sriburadej, Qi Li, Jeremy Espino, Ronald Voorhee, "Probabilistic Case Detection for Disease Surveillance Using Data in Electronic Medical Records", *Online Journal of Public Health Informatics*, Vol. 3, No3, 2011 (http://ojphi.org/article/view/3793/0)

[7] J. C. Roberts, "State of the Art: Coordinated & Multiple Views in Exploratory Visualization", *Coordinated and Multiple Views in Exploratory Visualization, 2007. CMV '07. Fifth International Conference on*, Zurich, 2007, pp. 61-71.

APPENDICES

Appendix A: Interview Findings Report

Observations and Semi-Structured Interviews

The first step in conducting user studies for this project was observations and semi-structured interviews with professional health care and public health workers. These interviews were intended to address the following technical objectives of the grant:

- Design: Elucidate design objectives
  - What are the additional needs that are unique to population health, of practicing health professionals, including providers, administrators and planners?"
- Data: Develop a scalable approach to deal with the sheer size and complexity of the data through the use of concise and controlled meta-data representation

Intent of Addressing Questions

- Discover tasks and questions/intents that can be addressed by CommonGround.
- Define scope of development, data requirements and ontology
- Inform application tasks and data relationships
- Help define scope of the usability studies.

Findings from the Interviews

The following sections are categorized by the data; how data are collected, used, processed and shared. Each section includes *intents* written from the user's perspective (e.g., "I want or need" or "I do…"). These intents are based on observations or statements made during the interviews. Some interviewee statements are included for each *intent* to provide context and insight.

Data come from different places in different ways…

Some data are reported to me

- "Most data come in from lab results"
- "Intermountain has data for all requested labs, but non-Intermountain only send positive lab results"
- "Data come from investigations conducted by local health departments"
- "Of the data that come into the state health department, only 30% are electronic. There are people whose full time job is to enter data into the system."
- "Death certificates give primary cause and contributing factors"
- The state DOH uses student absentee lists (all causes are lumped, non-specific)

Some data I pull from other sources

- NIS (National Inpatient Sample) and KID are given in some part free online

- “National hospital has data from 1992 – 2010. The quality of data changes, the fields change yearly based on research requests. However, some data are consistent”

When I get the data…

I combine data from multiple sources

- “It would be nice if you could ask the data one thing and get all the information instead of asking for this part and then this part and then another part.”
- “It is hard to compare data sources because most data sources are not compatible”

I fix inconsistencies

- “A lot of data inaccuracies are based on slight changes in the way the data was entered” (e.g., UVH vs. University Hospital) “
- The thing that I spend the most time on that I don’t think I should have to is the cleaning of the data; which can take three or four times as long as running an analysis”

I want to be efficient and leverage other people’s work

- “I don’t want to re-manipulate data that other people have already done”
- Every state can use their own reports, although many use similar or identical ones

When I consider the data…

I know that uncertainties are associated with data

I know i don’t have all the data i want/need

- “my biggest concern is that we just do not have enough data that are usable”
- “we can only make decisions based on the data we have available”
- “some data are so limited that it is useless (sporadic, small sample sizes, etc.)”
- “nobody knows if mdros stop becoming infectious over time”
- “sometimes it makes you wonder what is really out there”
- “we are creating a new form to get more information – ask ‘where will the patient be discharged to?’”

Some data requires me to gather more data

- “We get delays when we ask for special reports from hospitals”
- “When we have so much data about some diseases (e.g., hepatitis C or chlamydia) we don’t have time to trace each disease because there are just too many cases. If there was a program that allowed us to just grab the data, we could do so much more.”

Some sources and data types are more reliable than others

- “Some laboratory tests are better than others – each lab sometimes has different tests”
- “A lot of the data we use are reported voluntarily, this causes bizarre shifts because of the lack of reporting”
- “Vaccination information is difficult to collect and draw conclusions from – lots of inaccuracies”
- “Intermountain gives a lot of good digital data – regular and very complete; however, when matching data from separate report chains there are usually some missing data.”

Expertise and preferences affect how I interpret or comprehend the data

- “Before I start I like to look at the data and make sure it makes sense”

- "If the pathogen diagnoses are higher than average it doesn't necessarily mean an outbreak – it could be pathogen awareness or better reporting."
- "We take some liberties because this is a population that is used to seeing this data"
- "The numbers that I think are important are sometimes not what other people think are important"

I look at the data in different ways

I need to manipulate data in different ways

- "I would like to be better able to visualize data"
- "It is very hard to create a table that will portray this much data as quickly as this graph."
- "To manipulate the data I use R and Stata and SAS"
- "To use the data I pull it into SAS and sometime manipulate in Excel to create charts and graphs."

I make comparisons between groups (filter, sort and link)

- "You split up the age group based on how you want to look at the data."
- "Age groups change for each disease"
- "We match ICD9 codes to inpatient costs within hospital faculty to measure variability of cost and performance"

I make comparisons between time periods to interpret the current data and/or recognize when to take action

- "I need a large amount of information over a large number of patients to detect any patterns"
- "Deciding when diseases are considered epidemics/endemic is based on a numerical approach"
- "We define seasons for pathogens from mid-summer to mid-summer because that is when pathogen season is"
- "There are problems with using too many years' data, they can water it down and lose some relevance."
- "Data anomalies can cast shadows for years to come"
- "Thresholds for disease surveillance are based on 5-year trends (fairly stable from year to year)."
- We compare to averages of prior seasons"

Algorithms for analyzing the data can create misleading output

- "Bigger facilities will have more samples so that the confidence intervals are not the same – some places have more uncertainty"
- "I think that smaller facilities are shortchanged in reports. They cannot be above average on the safety listing because do not have enough line hours to be considered "green" and they never will be."
- "The statistical program we use builds on data during a 'training session.' Unfortunately, if you have an epidemic during the training period then the program sees it as 'endemic.'"
- "We always use at least two statistical models so that one can cover the other one's weaknesses."
- "One of the limitations of a statistical model is that if it is not 'significant' then it is off the radar"

The data affect my actions

I use data to predict future events

- "I create a model by iterating fit and comparing expected data with real data."
- "We are trying to forecast disease like the weather."
  - "Forecasting is generally not accurate."

Data trigger actions based on thresholds

- "Chicken pox is endemic but 5 or more cases in a school means that non-vaccinated people are excluded"
- "Based on the state definition 2 cases of pertussis in a public location is an outbreak."

Other people and organizations affect or dictate what I do…

I collaborate with others

- "The only way to work with local departments I through personal relationships"
- "Local DOHs will investigate possible outbreaks by contacting hospitals."
- "We create an initial report that goes around the office, then goes to the hospitals who have a chance to look over and request changes before we release it to the public"
- "There is a collaborative effort (60 people and 5 subcommittees) that follow MDROs. They are trying to create patient transfer guides."
- "All the data I get an epidemiologist somewhere here gets it too."
- In case of a novel flu strain, the state DOH will notify the CDC, World Health Organization (WHO), and work closely with local health departments.

I share with others

- "I get calls from the DOH with a request to look up specific pathogens"
- "Sometimes there are calls from doctors with personal queries or concerns"
- Local DOHs try to contain outbreaks by giving hospitals information packets
- The graphs are posted on the internet and emailed to local health departments during flu season
- We create weekly evaluation reports for all diseases, including flu reports and enteric diseases

Data need to be structured differently depending on the audience

- "Hospitals are interested in their own demographics and how they rank up."
- "<local health departments> are mostly interested in their own area."
- "'Joe Schmo' doesn't want to read 10 pages."
- "People in the hospital want more info than 'Joe Schmo'"
- "The state DOH releases information so that the public can understand it."

Sometimes I can't see or share data because of policy and security issues

- I would like to have data sharing and collaboration permissions
- "user agreement problems keep data out of reach."
- "…biggest concern is patient privacy"
- "Only parts of the DOH are allowed to get access to electronic health records."

My work is limited by the tools I have to use…

- "If you choose the wrong date format it messes up the whole thing" (referring to Pentaho).
- "We don't get the great tools that the U comes up with"
- "Public health is far behind the medical community as far as electronic exchange."

Appendix B: Database Schema

The data used by CommonGround is stored in a relational database using the following schema.

Encounter: Demographic information per encounter. The table does not depend on foreign keys and new encounters can be added at any time.

id	integer	pk
date	date	
age	integer	
zipcode	text	

KB: ontology of medical findings (tags). Appendix C describes the the ontology is detail. The ontology doesn't contain foreign keys and is thus independent of the rest of the data. The table can be augmented with additional finding definitions.

id	integer	pk
name	text	
category	text	
system	text	
details	text	

EncTag: a many-to-many association between encounters and findings

enc_id	integer	fk: Encounters.id
tag_id	integer	fk: KB.id

DetectorInfo: names of available detectors.

id	integer	pk
name	text	

Detectors: case detector probability per detector and encounter.

did	integer	fk: DetectorInfo.id
enc_id	integer	fk: Encounter.id
prob	real	

PathogeInfo: names of the various pathogens

id	integer	pk
name	text	
label	text	

Pathogens: positive identification of pathogens for each encounter

enc_id	integer	fk: Encounters.id
path_id	integer	fk: PathogenInfo.id
positive	boolean	

Appendix C: Knowledge Base

The CommonGround ontology consist of three classification groups (category, system, details) and 86 findings. Each finding is associated with one category and one system but only a few are associated with details.

Category	sign, symptom, syndrome, pathogen
System	respiratory, gastrointestinal, constitution, vision, other

Name	*Category*	*System*	*Details*
apnea	sign	constitutional	
cyanosis	sign	constitutional	
fever	sign	constitutional	low
fever	sign	constitutional	high
ill-appearing	sign	constitutional	
lymphadenopathy	sign	constitutional	
poor feeding	sign	constitutional	
rigor	sign	constitutional	
toxic appearance	sign	constitutional	
abdominal tenderness	sign	gastrointestinal	
seizure	sign	neurological	
chest wall retractions	sign	respiratory	
conjunctivitis	sign	respiratory	
conjunctivitis	sign	respiratory	
crackles	sign	respiratory	
dyspnea	sign	respiratory	
grunting	sign	respiratory	
hemoptysis	sign	respiratory	
hoarseness	sign	respiratory	
hypoxemia	sign	respiratory	
nasal flaring	sign	respiratory	
rales	sign	respiratory	
rhonchi	sign	respiratory	viral
staccato cough	sign	respiratory	
stridor	sign	respiratory	
tachypnea	sign	respiratory	
wheezing	sign	respiratory	
anorexia	symptom	constitutional	
arthralgia	symptom	constitutional	
chills	symptom	constitutional	
decreased activity	symptom	constitutional	
fatigue	symptom	constitutional	
malaise	symptom	constitutional	
myalgia	symptom	constitutional	
abdominal distress	symptom	gastrointestinal	
abdominal pain	symptom	gastrointestinal	
diarrhea	symptom	gastrointestinal	
nausea	symptom	gastrointestinal	
vomiting	symptom	gastrointestinal	
headache	symptom	neurological	
chest pain	symptom	respiratory	
cough	symptom	respiratory	productive

cough	symptom	respiratory	nonproductive
cough	symptom	respiratory	other
cough	symptom	respiratory	barking
respiratory distress	symptom	respiratory	
runny nose	symptom	respiratory	
sore throat	symptom	respiratory	
stuffy nose	symptom	respiratory	
bronchiolitis	syndrome	constitutional	
bronchitis	syndrome	constitutional	
croup	syndrome	constitutional	
ILI	syndrome	constitutional	
pneumonia	syndrome	respiratory	viral
pneumonia	syndrome	respiratory	other
CAMPYLOBACTER	pathogen	gastrointestinal	
CRYPTOSPORIDIUM	pathogen	gastrointestinal	
EHEC	pathogen	gastrointestinal	
GIARDIA_LAMBLIA	pathogen	gastrointestinal	
ROTAVIRUS	pathogen	gastrointestinal	
SALMONELLA	pathogen	gastrointestinal	
SHIGELLA	pathogen	gastrointestinal	
ADENOVIRUS	pathogen	respiratory	
BORDETELLA_PERTUSSIS	pathogen	respiratory	
CHLAMYDOPHILA_PNEUMONIAE	pathogen	respiratory	
CORONAVIRUS	pathogen	respiratory	OC43
CORONAVIRUS	pathogen	respiratory	229E
CORONAVIRUS	pathogen	respiratory	HKU1
CORONAVIRUS_NL63	pathogen	respiratory	
ENTEROVIRUS	pathogen	respiratory	
H1N1	pathogen	respiratory	2009
INFLUENZA	pathogen	respiratory	B
INFLUENZA	pathogen	respiratory	A
INFLUENZA	pathogen	respiratory	NOT_TYPED
METAPNEUMOVIRUS	pathogen	respiratory	
MYCOPLASMA_PNEUMONIA	pathogen	respiratory	
PARAINFLUENZA	pathogen	respiratory	2
PARAINFLUENZA	pathogen	respiratory	4
PARAINFLUENZA	pathogen	respiratory	3
PARAINFLUENZA	pathogen	respiratory	1
PARAINFLUENZA	pathogen	respiratory	
RHINOVIRUS	pathogen	respiratory	
RSV	pathogen	respiratory	
SEASONAL_H1	pathogen	respiratory	
SEASONAL_H3	pathogen	respiratory	
SEASONAL_NOT_TYPED	pathogen	respiratory	