AD_____

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:
April 2014

TYPE OF REPORT:
Annual

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

DISTRIBUTION STATEMENT:

x Approved for public release; distribution unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
April 2014	Annual	5 March 2013 – 4 March 2014

4. TITLE AND SUBTITLE
Common Ground: An Interactive Visual Exploration and Discovery for Complex Health Data or Complex Health Data

5a. CONTRACT NUMBER
W81XWH-13-1-0028

5b. GRANT NUMBER
W81XWH-13-1-0028

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Yarden Livnat, Per Gesteland, Adi Gundlapalli

E-Mail: yarden@sci.utah.edu, per.gesteland@hsc.utah.edu, Adi.Gundlapalli@hsc.utah.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

UNIVERSITY OF UTAH
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 to enhance situational awareness by providing an effective visualization of large, complex and heterogeneous population health data. Presently, users of complex health data are overwhelmed with charts, graphs, tables and maps. Our goal is to develop a novel health data weather map visualization prototype that provides a dynamic, interactive presentation of complex healthcare data. We received IRB approvals from TATRC, University of Utah and Intermountain Healthcare in September of 2013. We signed a data share agreement with Intermountain Healthcare and expect to receive initial limited dataset by May 2014. To elucidated design objectives and inform the visual interface design, we use methods of cognitive task analysis and conducted structured observations and contextual interviews with practicing, front-line public health epidemiologists working at the State and local health departments in Utah. We developed a graph-based database that is specifically tailored for this project and installed an intermediate dataset that represent the data we expect to get from Intermountain Healthcare. We began work on a special purpose ontology that will be used to describe data and annotate other ontologies for the visual interface client. Finally, we are actively working on software development of both a backend server and the advance web-based visual analytics display.

15. SUBJECT TERMS
Visualization, Visual Analytics, Ontology, Software, Population Health Data

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

TABLE OF CONTENTS

1 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]. We hypothesize that a well-designed visualization interface that is tailored to the users 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.

2 Body

2.1 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 practicing, front-line public health epidemiologists working at the State and local health departments in Utah. These interviews were intended to address the following technical objectives of the grant:

- Section 3.1 - Design: Elucidate design objectives.
  What are the additional needs that are unique to population health, of practicing health professionals, including providers, administrators and planners?
- Section 3.2 - 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

The interviews addressed the questions of,

- 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.

We have conducted six contextual interviews in addition to the original two pilot interviews. The participants included two epidemiologists and an analyst from state health department, an epidemiologist and analysts from a county health department, and a research pediatrician. We had planned and had initial contact with Joe Lombardo from the biosurveillance group at the Johns Hopkins University Applied Physics Laboratory (APL) to conduct contextual interview at their laboratory. APL has been developing the Essence and Essence II Syndromic surveillance systems and would have been an exceptional source of valuable insight. However, we were unable to schedule these interviews as planned. A five-page report on the interview findings with concept categories and evidence to support intent is provided in Appendix A.

The CommonGround visual paradigm presents a novel and unique user interface and interactions. An earlier software prototype (EpiCanvas) was been well received [2][3]. As part of the design phase in this work we aim to better understand how natural and intuitive this approach is and evaluate various aspects of it. In particular, we aim to understand and evaluate how users

- Perceive and understand the arrangement of information (i.e. conceptual tags representing meta-data) on the screen,
- Comprehend the dynamic visualization of relationships between such tags
- Are able to identify patterns in the data
- Comprehend changes over time and over space.

To evaluate these questions we elected to conduct a user study using undergraduate students from the department of Psychology at the University of Utah. We have worked with the Department of Psychology to enable us to conduct these user studies in their lab. The pool of highly trained public health officials in SLC that could be available for our studies is limited and we opted to enlist their help for the software prototype evaluation study at the end of this project. We have been working on the design of the user study with the students, identifying the questions, the data for the study, and the changes that need to be incorporated into our prototype in order to collect the appropriate information. The user studies will take place in Spring 2014.

2.2 Data

2.2.1 Scalable data repository

Both the design and the software development phases in this project rely on 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 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.

The second source for data that we identified in the proposal is Intermountain Healthcare. The advantage of these data is that they are 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 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 worked with Intermountain Healthcare, the University of Utah IRB and TATRC IRB in order to get permission to use such data. The need for multiple agencies granting approval resulted in delays. We received IRB approval from Intermountain Healthcare and signed a Data Use agreement with Intermountain Healthcare on May 14th 2013. We also received IRB approval from the University of Utah on June 5th 2013. However, we received a final IRB approval from TATRC only on September 10th 2013, more than six months after the start of the project. We have been working with Intermountain Healthcare to obtain data (awaited as of writing of this report). from the key issue is the HIPAA related requirement that Intermountain Healthcare will clean, transform and annotate the data before the data are given to us. This phase proved to be a challenge yet we anticipate we will begin to receive data in May-June 2014.

2.2.2 Ontology

Focusing on the development of a pilot ontology to capture relevant knowledge for population health data, we developed and installed the following infrastructure and resources.

For the development and management of the ontologies, we installed a framework consisting of a server (Apache Tomcat), a triple store (Sesame by Aduna), a reasoning engine (Pellet by Clark & Parsia), and several development tools (Protégé OWL from Stanford University with OWLViz, and SHRIMP by Harvard University). The server hosted the various web-based applications and tools listed above. Ontology concepts and instances were stored in the Sesame triple store, and accessed with the included Sesame query and management tool. Ontology analysis and testing was realized with the Pellet reasoning engine. Finally, Protégé was the main ontology development tool, with ontology visualization provided by OWLViz. SHRIMP (Simple Ontology Mapping Tool) was installed for enriching Protégé with ontology mapping functionalities. For ontology representation, we selected the OWL 2 RL standard profile its good expressive power, scalable reasoning capabilities, and history of successful applications in the biomedical and public health domains

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.

2.3 Presentation

2.3.1 Software Prototype

The original software tool was developed as a small desktop application using Adobe Flex framework (http://www.adobe.com/products/flex.html). Our aim in this project is to develop a web-based application that can easily be access from any modern web browser. To achieve this goal our new design is based on 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. This in turn means that the client does not depend on a specific type or implementation details of the data repository. It also reduces security concerns as the data repository is kept behind a firewall and is accessible only by the dedicated server. The new architecture comprises of three layers: the data layer, middleware and a presentation layer.

2.3.2 Data layer

For the development of a versatile data layer we considered various types of data repositories. The main features we considered were scalability, ease and extensibility of the query mechanism, the ability to easily integrate ontology data, and the capability to annotate the data with ontology based information.

Relational databases are ubiquitous and we have used an embedded SQLite (https://sqlite.org/) in the original tool. Data access using SQL is quite powerful as well. However, both relational databases and the SQL query language are not well suitable for representing hierarchical data such as ontology. Scalability can also pose an issue it is one of the reasons large-scale corporations have migrated in recent past to key-value based repositories such as Apache Hadoop (http://hadoop.apache.org). Key-value based data storage is more appropriate for representing ontologies and annotated data than relational databases and can scale better in certain situations. However simple key-value data repositories provide too low-level access and structure and impose on the user the need to implement all the higher-level structure and access. We also experimented with RDF based repositories such as Sesame (http://www.openrdf.org/). Resource Description Framework (http://www.w3.org/RDF/) is a standard model for data interchange on the web and is well suitable for storing ontologies. There are various implementations of RDF based repositories both proprietary and open source. Much like the Hadoop type of storage, RDF based systems provide very little support for complex data representation are require that all data representation and access be in form of triplets.

The best solution we identified for the purpose of this project is a graph based data repository.. In particular we chose the very successful Neo4j implementation (http://www.neo4j.org/). Neo4J is a highly scalable native graph database that is widely used in both commercial and government's projects. Graph databases are well suitable for representing large heterogeneous data, storing ontologies and for annotating such data with ontology-based information. We deployed a graph database

repository on sever at the SCI institute and designed a data schema appropriate for this project. As discussed above we expect to receive data from Intermountain Healthcare in early May. In the meantime we converted and loaded a small relational dataset we used for the original tool into the newly created graph-based data repository.

Security and HIPAA compliance area critical issues that must be addressed when dealing with protected health data. The Center for High Performance Computing (CHPC) at University of Utah (http://www.chpc.utah.edu/) has recently created a HIPAA compliant protected sandbox environment (https://www.chpc.utah.edu/docs/research/featured/CHPC/The+HIPAA+Cluster+-+Ensuring+Data. We are working with the CHPC to create a development setup within this environment that includes both the graph-based data repository and a server. The new setup will ensure that only authenticated users using our client will be able to access our data and only through this server.

2.3.3 Middleware

The middleware consists of a server that facilitates communication between the web-client and the data repository. The main role of the server is to form a security barrier between the outside world the data repository and thus reduce the risk of an unauthorized access to the limited dataset through exploits of security flaws, much the same way as SQL injection (http://en.wikipedia.org/wiki/SQL_injection). The server is also used to reduce the amount of data the must be send to the client and to provide a scalable platform for performing complex and computational intensive operations on behalf of the client.

W have experimented with Apache Tomcat (http://tomcat.apache.org/), which is a Java based server technology. As described in the following presentation layer section we initially planned on using a java-based client and thus having a java-based server seemed a logical solution from a development perspective. There is no fundamental need to develop both the client and the server using the same language. One of the issues we want to reduce the development and maintenance costs and having the consistent environment for both the client and the server help simplify them. Tomcat also requires a special installation that we wish to avoid.

After considering and evaluating various options we elected to develop our server based on NodeJS technology (http://nodejs.org/). NodeJS a platform based on Chrome's JavaScript runtime that is very lightweight and efficient making it perfect of data-intensive real-time applications. An additional advantage of using a JavaScript based server is the native support for JSON (http://www.json.org/) based data-interchange format that greatly simplify the communication between the server and the web-based client.

2.3.4 User interface

The development of the user interface software was delayed until the start of the academic year when we were able to recruit the two computer science students. The original prototype we developed as part of a previous TATRC funded effort was based Adobe Flex software framework. This technology is being subsumed by JavaScript and HTML5 technologies and will not be supported for much longer. We have experimented with alternative frameworks and initially selected to develop the new prototype using the Vaadin framework. Further work with Vaadin has proved the framework is not adequate for our needs.

We have since moved to develop the software using JavaScript/HTML 5 technologies and in particular the AngularJS (http://angularjs.org) framework that is being developed by Google. Using these

technologies, we developed an HTML5 client that runs on Windows, Mac OSX, Linux and mobile systems such as the iPad.

The visual interface software is in an active development phase. We developed the underlying software framework of the web-client that supports communication with the backend server and visualizing the data. The main display features a dynamic graph view of tags in a form of a tag cloud and is built on top the D3 graphics library (http://d3js.org). The size of each tag represents the number of reported cases associated with that tag. We've implemented three different layout algorithms to spatially arrange the retrieved tags on the screen. During an investigation a user can switch between these layout algorithms and any additional methods we will develop a part of this project. The simplest algorithm features a random layout while a second iterative algorithm employs a well-known graph layout technique. We are actively working on developing an advanced third layout algorithm that incorporates more information about the relationship between the tags (e.g. number of shared reported cases). The algorithm also aims to avoid and remove any visual overlap between the tags.

The web-client also features a map view using the Leaflet (http://leafletjs.com) open software JavaScript library. The map view enables the user to see the spatial distribution of various subsets of the reported cases aggregated based on zip codes. The display of the zip codes layer poses a challenge when a user zooms in and out the map. To facilitate fast zooming capabilities we store several zip code shapefiles (http://en.wikipedia.org/wiki/Shapefile) at different resolution on our server and dynamically fetch the most appropriate one at run time. We incorporate charting capabilities into the software to visualize bar- and line charts.

The software development of the visual display and the backend server are currently the main focus areas.

2.4 Evaluation

The project has not reached the evaluation phase yet. We will enlist public health personnel to evaluate the final software prototype.

2.5 Problems/Issues

The IRB process had taken over 6 months, much longer than expected. We received the final IRB approval on September 19th 2013. This introduced a long delay in conducting user studies and contextual interviews, which in turn hamper our effort to understand the needs and the development of novel visual representations to address these needs.

The award was established only in March of 2013. Because the project started so late in the academic year (midway through the spring semester) we had initial difficulties in recruiting the two computer science graduate students. In September of 2013 we recruited two new master students and a third master student in January of 2014. We are working with the students to get them familiar with the domain of Information Visualization, the tools and various advance algorithms we are using in this project.

We have requested and received a no-cost extension until March 4th 2015 to complete the objectives of this project.

3 Key Research Accomplishments

- Received IRB approvals from Intermountain Healthcare, University of Utah and TATRC
- Conducted six contextual interviews with public health officials and authored a five-page findings report.
- Created a private database with sample population health data.
- Developed an initial client-server software prototype.
- Published and presented a full paper in the Workshop on Signature Discovery.
- Gave a presentation at Pacific Northwest National Lab.
- Awarded an NSF SBIR grant ($150K/6 months)
- Submitted an NSF SBIR grant ($750K/2 years) – reached final review panel and awaiting for final funding decision.

4 Reportable Outcome

4.1 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.

4.2 Presentations

Y. Livnat, The CommonGround Visual Paradigm for Biosurveillance

- Pacific Northwest National Lab (PNNL), 2013
- Workshop on Signature Discovery, Intelligence and Security Information, Seattle, 2013

4.3 Informatics

Created a graph database (using Neo4J) that includes a limited dataset about pediatric patients presenting to an Emergency Department (ED) in Salt Lake City between 7/2007 and 6/2008. The data types available for these patients include demographic information, chief complaints, syndromes and specific infectious germs. The database does not contain personal identification data and is defined as a limited dataset only because it includes dates and zip codes information. The database is only available for and with in the development phase and it not publically available.

4.4 Funding based on this work

The following are NSF funding based on the visualization we developed as part of work. The proposals demonstrate the generality of this visual paradigm to wide range of other domains.

1. 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

2. NSF SBIR Phase II (submitted)
   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
   Status: Reached final panel; awaiting final decision as of April 15, 2014

5 Conclusion

In summary, we have made significant progress on the objectives of this project. With the goal of maturing a prototype for enhancing situational awareness by effective visualization of large, complex and heterogeneous population health data, we have (1) elucidated appropriate design objectives by conducting contextual interviews; (2) made progress in obtaining healthcare data sets from a large operational partner (Intermountain Healthcare); (3) developed an ontology to be used for this project and (4) made progress on the software prototype. We have experienced delays with regard to regulatory approvals and access to healthcare data sets that are beyond control. During the no-cost extension period, we will complete the objectives of this project, perform an evaluation of the software and submit a final project report.

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 clinical domains. Our results and deliverables 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

6 Reference

- 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.
- 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
- 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)**

7 Appendices

- Appendix A: Observation Interviews Initial Findings

CommonGround – TATRC
Observation Interviews Initial Findings

Date	Author	Comments
12/16/13	Heidi Kramer	Based on the first set of user interviews
1/6-9/14	Heidi Kramer	Added categories and details

Document Purpose and Background

This doc contains interview findings with concept categories, and evidence to support intents.

Next steps: use the data from this document to create requirements and design docs for CommonGround.

1. Initial 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:

"3.1 Design: Elucidate design objectives

- What are the additional needs that are unique to population health, of practicing health professionals, including providers, administrators and planners?"

"3.2 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 data 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 make 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 deparments.

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."