Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Sorie Dumbuya	Date

Applying Public Health Informatics to Developing a Training Database System and Investigating an Influenza Outbreak

By

Sorie Dumbuya
Master of Public Health

Applied Public Health Informatics

_____
Committee Chair

_____
Committee Member

Applying Public Health Informatics to Developing a Training Database System and Investigating an Influenza Outbreak

by

Sorie Dumbuya

B.S., Georgia State University, 2009

An abstract of
A thesis portfolio submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Applied Public Health Informatics
2017

Abstract

Applied Public Health Informatics to Developing a Training Management Database and Investigating an Influenza Outbreak

By Sorie Dumbuya

Background: Public health informatics (PHI), the science of managing public health information, is a relatively new field that encompasses wide-range of disciplines. The goal of PHI is to find the best way to apply technology to support the work of other public health professionals who use health information and management systems. The following portfolio comprises the planning and design of two information systems proposed to make information available to stakeholders, allowing me to apply my knowledge, skills, and abilities in public health informatics.

Key Aims: The goal of the first project was to plan and design a database system for National Public Health Institutes to increase their ability to efficiently operationalize, manage, monitor and evaluate their training efforts. The goal of the second project was to design an information system that would be used to facilitate the outbreak investigation and surveillance of a hypothetical emerging influenza in low-resource settings.

Methods and Results: A detailed design of the database system was developed by conducting all of the tasks for the early phases of the Database System Development Life Cycle. A high-level design was developed for the influenza information system based on similar systems and leveraging existing technologies.

Conclusion: While much work remains, I was able to both utilize my experience and gain new knowledge by laying the foundation for the establishment of both systems. With thorough, appropriate planning, each system would have a significant impact on public health if developed.

Applied Public Health Informatics Thesis Portfolio

by

Sorie Dumbuya

B.S., Georgia State University, 2009

An abstract of
A thesis portfolio submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Applied Public Health Informatics
2017

Table of Contents

Introduction

The digital revolution, marked by the move from mechanical and analog electronics to digital electronics, has catapulted the world into the age of information. The Information Age is characterized and aided by the rapid increase in human ability to store, share, and compute digital information (Hilbert & López, 2011). This has led to an era of open information known as the big data revolution, named so for the volume, complexity, timeliness, and diversity of data being collected across industries (Groves, Kayyali, Knott, & Van Kuiken, 2013).

In order to prepare for this data revolution, public health and healthcare organizations across the globe are launching initiatives to improve the quality and increase the amount of information available. As a result, data is becoming more accessible, usable, searchable, and actionable for the healthcare (Groves et al., 2013).

All of these initiatives are enabled by the growing field of public health informatics (PHI). As the Framework Institute defines it, public health informatics is the science of managing public health information, where informaticians "ensure that data are shareable and design and implement the integrated systems for sharing health data that are crucial to public health practice and high-level decision-making" (Fond, Volmert, & Kendall-Taylor, 2015).

As a public health informatician, my role is to figure out the best way to apply technology to support the work of other public health professionals who use health information and management systems to collect, store, retrieve, analyze, interpret, present, and disseminate health information to improve population health outcomes (Whittaker, Hodge, Mares, & Rodney,

2015). With this in mind, two systems that would make information available to stakeholders were planned and designed in order to improve health outcomes.

Working on these projects allowed me to apply knowledge, skills, and abilities I have gained through the program. The specific knowledge and skills in which I gained the most experience from working on these projects were project management, stakeholder analysis, requirements gathering, database design, and epidemiology.

Overall, I had the opportunity to consider the mission, goal, business needs, workflow and information flow of an organization or team to develop a list of requirements. Based on these requirements, I had to define system specifications, diagram models, and design information systems that would help stakeholders the meet their information and data needs.

National Public Health Institute Training Database System

Introduction

Problem Statement

National Public Health Institutes (NPHI) are science-based governmental organizations that serve as the focal point for public health efforts of their respective country. NPHIs are being stood up all over the world, particularly in developing countries, where Ministries of Health, and governments as a whole, face many challenges with providing basic services to their people. While Ministries of Health working tirelessly to provide and regulate healthcare delivery, they are left with minimal resources to also execute public health initiatives. NPHIs aim to fill this void and conduct many activities to recognize, anticipate, and respond to needs or problems in support of Ministries of Health. In addition to surveillance, research, and policy development, training is one of these activities NPHIs will conduct to ensure improved health outcomes.

NPHIs train hundreds of individuals in various public-health related skills and competencies each year. These include, but are not limited to, epidemiology, project management, research, community outreach. With minimum resources and formidable challenges, NPHIs must conduct their training effectively and efficiently. Therefore, sound monitoring and evaluation must be in place.

NPHIs will need to manage information for trainees currently enrolled in training, graduates, trainers, and the various training courses to not only efficiently manage enrollment but also monitor and evaluate the progress and impact of the program. Developing a database will

enable NPHI administrators to store and retrieve information in an accurate and effective manner.

An NPHI database system would have a significant impact on their business processes. First and foremost, training Information will be more accessible to team members, trainees, alumni, and other individuals associated with the institute as the database will be accessible via the web. In addition, since forms can be sent out with reminders and notifications, more data can be easily collected. Another benefit the database will provide to an NPHI is that it will save resources, namely time. With just a few clicks in the computerized database, retrieving a single record that may have taken several minutes to locate in a paper-based filing system can be done instantly. Also, reporting would no longer have to be done manually, eliminating hours of retyping. NPHI has limited resources and implementing this database will free up man hours and office supplies.

Purpose Statement

The purpose of this project is to provide a high-level description of a database system and the project plan to implement it that could eventually be used by an NPHI. This project strategically plans the design and implementation of a digital, online, cloud-based database system to improve an NPHI's training efforts and enhance its overall contribution to public health in its country.

Methods

The NPHI training database system must be developed in a way that it not only fits within the NPHI's current activities, but it also can contribute to the NPHI's progression and

advancement. This project goes through the database system development life cycle for a NPHI Workforce Development Training Database. The database system development life cycle consists of: requirements analysis, database design, evolution and selection, implementation, data loading, testing and performance tuning, operation and maintenance, and growth and change (OpenLearn, 2016). The primary activities for the requirements analysis and database design phases of the database system development life cycle were done for this project.

Requirements Analysis

Organizational Analysis

The successful establishment of the database system must align with the mission, goals, and processes of the NPHI. Thus, the essence of the NPHI, its mission to improve the health of the people, its strategy for achieving this mission, and its organizational structure should be taken into consideration during the entire planning of the database system.

"NPHIs focus on the major public health problems affecting the country. They use scientific evidence as the basis for policy implementation and resource allocation and are accountable to national governments and the public. Their key functions—including disease surveillance, detection, and monitoring; outbreak investigation and control; health information analysis for policy development; research; training; health promotion and health education; and laboratory science—are particularly critical in low-resource nations."

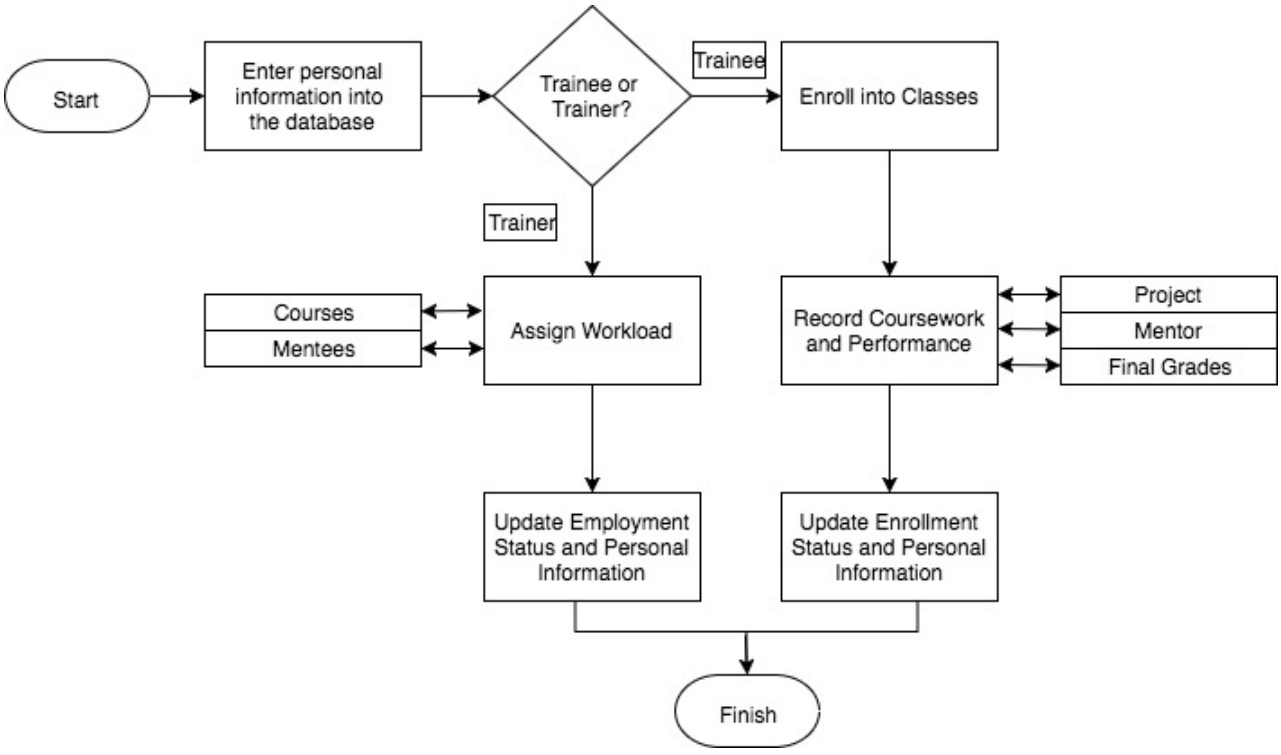
(International Association of National Public Health Institutes, 2017)

NPHIs improve the capacity of public health professionals through training as one of its key functions. The database system will be stood up to manage the information related to the

training. This includes information about the trainees, trainers, training course, and alumni. Figure 1. displays the workflow process currently used by the Liberia Field Epidemiology Training Program (FETP), which an NPHI would likely use.

Using the Liberia FETP as a model and meeting with their representatives to understand their training goals, potential requirements for the database were determined. This includes, but is not limited to, what data and how much need to be collected, what technologies must be used, and what reports and visualizations were needed.

Figure 1. Example National Public Health Institute Training Program Information Collection Workflow Process



Database System Requirements Definition

The business, functional and non-functional requirements are effectively defined below. Defining the database system requirements is the primary objective of the requirements analysis phase of the database system life cycle. The intended audience for this list of requirements is the project manager, project team, project sponsor, client/user, and any stakeholder whose input/approval into the requirements definitions process is needed.

Business Requirements

1. The NPHI training program must improve the capacity of the public health workforce.
2. The NPHI training program must be flexible enough to support current standards, but also adapt to new standards.
3. The NPHI training program must position itself to support the National Ministry of Health and other public health organizations.
4. The NPHI training program should be able to collaborate with other NPHIs.
    1. There is no funding for additional Full Time Employees.

Functional Requirements

Trainee Management

1. Personal information (e.g. name, address, contact information, date of birth, educational background, work experience) about the trainees enrolled in the program must be captured and stored in the system.
2. Information about the trainees' enrollment (e.g. courses taken, period and duration of training, level) must be captured and stored in the system.

3. Information about the trainees' course work and performance (e.g. projects, final grades) must be captured and stored in the system.
4. The system must be able to keep track of the trainee as their status changes.

Trainer Management

1. Personal information (e.g. name, address, contact information, date of birth, educational background, work experience) about the trainers must be captured and stored in the system.
2. Information about the trainers' employment and assignments (e.g. courses taught, period and duration of employment, mentorship) must be captured and stored in the system.

Curriculum Management

1. Information about the courses (e.g. description, enrollment, location, texts) must be captured and stored in the system.

Alumni Management

1. Information about the graduates (e.g. name, address, contact information, date of birth, educational background, work experience, current employment) must be captured and stored in the system and updated regularly.

Reporting

1. Canned and ad-hoc reports must be supported by the system.

Non-Functional Requirements

Usability Requirements

1. The user interface must be user-friendly.
2. The system must be cloud-based.
3. The system must be accessible by personal digital assistants, tablets, and smart phones supporting multiple common operating systems for mobile devices.

Performance Requirements

1. The system must be scalable to support new individuals and courses.
2. The system must be available 24/7/365 with a maximum of 14 days of downtime annually for maintenance.

Supportability Requirements

1. The system must include on-line help.

Security Requirements

1. Personal privacy must be maintained since the system will contain personally identifiable information.
2. Log in may require two-factor authentication.
3. Sensitive data, such as passwords and credit card numbers, stored in the system will be encrypted.
4. The system will have firewalls.

Transition Requirements

1. The new system will be run in parallel with the old system for a period of 3 moths.
2. Employees for the NPHI training program will go through two separate full day training workshops before the system goes live.

3. Conference calls will be held to provide support for three months after the system goes live.
4. Data from the old system will be copied to the new system.
5. Documentation must be available, as well as a strategy to enroll new trainees, trainers, and courses.

Database Design

Specifications

After gathering requirements for the database, the system specification begins, resulting in design documents and a detailed description of how the system should be built. data models were drafted to determine the structure of a database and in which manner the NPHI data will be stored, organized and manipulated.

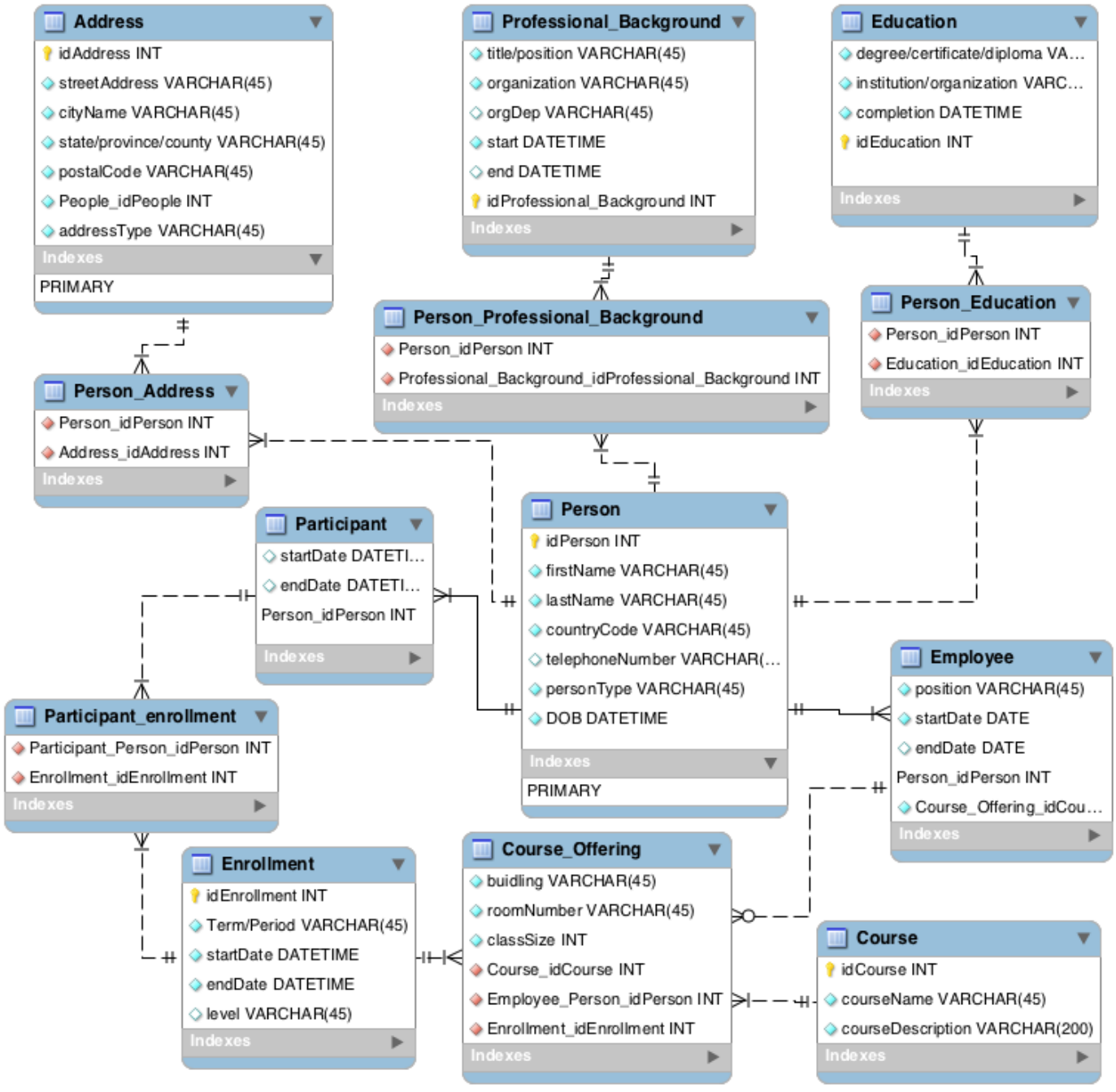
The first step in designing the database system is the conceptual data model, which specifies a logical schema. This schema determines the type of database that is required.

The database will relational, meaning the information will be presented in tables with rows and columns. The relational database management system will use SQL for many reasons. For one, standard SQL makes it is easier to manage the database system without having to write a great deal of code. Secondly, its established use cases and large community makes tech support more available and affordable. The database will be located in the cloud to take advantage of the security, reliability, availability, affordability, and computing power offered by many of the major vendors.

Logical Database Model

A logical database model, shown in Figure 2., was developed for the NPHI Training Database. It represents the definition, characteristics, and relationships of data, independent of a physical data storage device. (MicroStrategy, 2017)

Figure 2. Example National Public Health Institute Training Program Logical Database Model



Emerging Influenza Outbreak Investigation and Surveillance System

Introduction

The Emerging Influenza Outbreak Investigation and Surveillance System (FluSurv) will be a web-based information system that brings surveillance officers, epidemiologists, healthcare providers, lab technicians and the US Centers for Disease Control and Prevention (CDC) together to combat and monitor the spread of novel influenzas. It is being stood up for the specific outbreak going on in the Caribbean and Central America and will remain in place for future outbreaks. Due to the novelty and virulence of this strain of influenza, the outbreak in the Caribbean and Central America is a serious situation and must me controlled as quickly as possible. To this end, thorough outbreak investigation must be conducted and an effective surveillance system must be established. Meeting both of these objectives requires sound data collection, storage, and sharing. And the remote environment necessitates the investigation employ lower-power and offline approaches for the before mentioned tasks.

FluSurv Overview

This section provides a high-level description of the FluSurv system including its definition, primary business goal, business objectives, context, and capabilities.

Definition

FluSurv will be a web-based information system that brings surveillance officers, epidemiologists, healthcare providers, lab technicians and the US Centers for Disease Control and Prevention (CDC) together to combat and monitor the spread of novel influenzas.

Business Goal

The business goal of FluSurv is to take advantage of the Internet and World Wide Web to radically improve the way professionals investigate outbreaks and surveil the spread of novel influenzas.

Business Objectives

The business objectives FluSurv are to provide the following business benefits to surveillance officers, epidemiologists, healthcare providers, lab technicians and the CDC.

- Provide an interface and database for storage of information, such as line lists
- Show a shared map of places that have already been covered by investigation and places that still need to be investigated
- Calculate and display epi data, such as attack ratios and epi curves

External Hardware

FluSurv interacts, either directly or indirectly, with the following significant external hardware:

Client Hardware

- **User Client**, which are the personal computers, personal digital assistants (PDAs), and smart phones used by team members to communicate with each other and perform their tasks.

Networks

- **Internet**, which is the global network used for communication among users and FluSurv.
- **Local Area Network**, which is the FluSurv-internal local area network used for communication among users with limited internet coverage in the affected region and FluSurv.

Summary of System Capabilities

FluSurv will provide users with the following capabilities:

- Generate reports
- Search for individuals or locations
- Store case information
- Gather information for case-control studies
- Delete information
- Chat with other users
- Forecast the spread of influenza
- Review and manage their user information

Data Requirements

While all of the data being collected is important in one way or the other, they are not needed by everyone involved at every moment. In order to make things easier for ourselves and our collaborators, reduce the amount of data that needs to be transported, and protect the personal information of cases and contacts, certain data will be stored and transferred separately. The data that will be collected can be divided into three categories: cases and contacts, surveillance and epidemiology, and project management. Below is a brief description of each of the three categories.

Surveillance and Epidemiology

For every outbreak investigation, information must be gathered about each case. This process known as line listing involves collecting detailed information about each person suspected of being affected, such as demographics, exposure, timeline of events, etc. Another primary component of an investigation for an infectious disease is contact tracing, where individuals who have come in contact with suspected cases are interviewed, documented, and potentially monitored. In addition to information about the cases and contacts, other information (i.e. lab reports) to understand the prevalence, incidence, etc. fall into this category.

Cases and Contacts

The surveillance/epi data will be deidentified and the personal identifying information (names, address, contact information) will be stored separately. There will be a person id number used to match observations in the two data sets.

Project Management

In addition to data about the outbreak, data about the investigation itself will be collected and managed, as well. This will ensure that the investigation is conducted thoroughly and efficiently. It will also allow for lessons to be learned from the investigation that can be used for future reference. To this end, data regarding the individuals conducting the outbreak, the area covered/places visited, expenses, timeline of events, tasks executed, etc. will be collected.

Information Flow and Workflow Process

Being that this is an outbreak the system has to be up and running relatively quickly. To this end, I will launch a plan that will have the system up and data being transferred within three days of getting on the ground. The data for the system will flow to and from four stakeholders, the outbreak investigation and surveillance team, the healthcare facility, the lab, and the US Centers for Disease Prevention and Control CDC).

The final destination of the data will be a central repository on the cloud. As far as getting the data to this repository goes, there are a few options being considered. The general approach will be that the data from the various locations will be consolidated, separated into the various categories, and uploaded to the cloud.

Outbreak Investigation and Surveillance Team

Since the purpose of this project is to both investigate and control this outbreak and set up a surveillance system, there will be active and passive surveillance. The outbreak investigation and active surveillance activities will be carried out by this team. They will go into the

community and speak with suspected cases, their contacts, and community healthcare workers to get information to create line lists and potential sources of infection. They will also collect specimens to be tested at the lab.

The outbreak investigation and surveillance team will be able to access the system to view lab results and notes about all the cases. They are responsible for informing cases about positive or negative test results. They will also receive reports and visualizations about aggregate data, such as epi curves and heat maps.

Healthcare Facilities

Passive surveillance will take place at the healthcare facilities within and surrounding the affected area. When cases present to the facility, their information will be recorded and entered into the system. Healthcare providers at the facility will also conduct point of care testing, as well as collect specimens to be tested at the lab.

Healthcare providers will be able to access the system to view lab results and notes about all the cases. They will be responsible for informing patients about positive or negative test results. They will also receive reports and visualizations about aggregate data, such as epi curves and heat maps. They will also receive information about suggested treatment.

There will be a mixed-method of data collection. Data collected by the outbreak investigation and surveillance team will be primarily collected on mobile devices (phones and tablets) using digital forms and stored locally on the device. There will also be paper-based forms in the event that the devices run out of battery power. Data collected by the healthcare

facility will depend on the resources at each individual facility. If there is no internet, electricity, or computers at the facility, offline and/or paper-based methods.

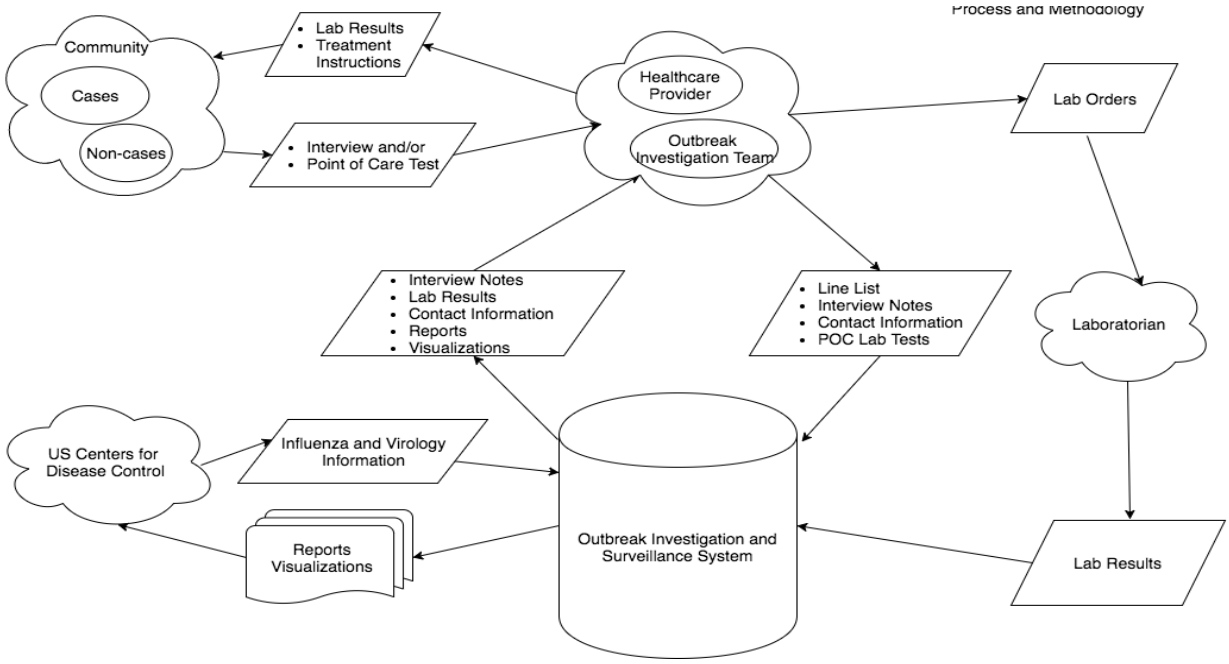
Laboratory

The lab will receive and test specimens. The laboratorian will enter this data, as well as information about the virus itself to the system.

CDC

CDC will have access to all of the de-identified data, as well as the lab results. They will also receive reports and visualizations about aggregate data, such as epi curves and heat maps. CDC will be responsible for entering information about similar or concurrent outbreaks in other regions. They will also suggest treatment options for patients.

Figure 3. FluSurv Information Flow Diagram



Application Design

The deployment of FluSurv will be done via Puppet, as it will allow us to define the application and its infrastructure. It will coordinate the spinning up of the database service, web servers, and load balancers needed by the application. With Puppet, any changes made on the master-server are propagated to all configured machines. Lastly, Puppet will facilitate collaboration between the team on the ground and the team at CDC to connect the application being stood up with existing information systems.

Table 1. FluSurv Application Components

Components	**Data Sources**	**Develop textual discussions and descriptions of the fit and function of each item on your design list.**
User-interface / Web server	Interviews / Case Reports / Lab results	The form is modular in design, collecting individual patient demographic information as well as relevant clinical, laboratory and prescribing information depending on user input. Each form is programmed for the Open Data Kit (ODK; www.opendatakit.org). The application's web server will be used to store, process and deliver the web pages that the different actors will use to enter data into the system.
Database	Interviews / Case Reports / Lab results	The database will be where the data collected and entered by the actors is stored.
Database Management System	Interviews / Case Reports / Lab results	The database management system will be used to create and manage the system. It will allow end users to create, read, update and delete data in the database.
Business Intelligence Layer	Interviews / Case Reports / Lab results and data from CDC	The business intelligence layer will sit on top of the database and provide dashboards for analytics, visualizations, and reports.

Data Analytics and Visualization

Specific Visualizations

FluSurv users will be able to create, display, and share many different kinds of visualizations. The analytics platform Looker™ will be added to the system to allow the team to be able to answer the questions they need to ask. Below are descriptions of the two key visualizations that will be available to users, epi curves and maps.

Epi Curve

An epi curve is a graph displaying of the onset of illness among cases associated with an outbreak. Epi curves are used to learn about the outbreak's distribution of cases over time (also known as the time trend), any cases that stand apart from the overall patter, the general sense of the outbreak's magnitude, the outbreak's pattern of spread, and the most likely time period of exposure.

Dot Maps and Choropleths

Maps will be useful to the team to determine the geographic distribution of cases. This information can help them determine the source of the outbreak, the spread of the outbreak, and the areas hardest hit. Dot maps are maps that show every case based on geography, irrespective of administrative boundaries, while choropleths show data aggregated based on administrative boundaries (see examples below). The dot maps will be particularly useful for those on the ground, while the choropleth will be useful to show politicians for resource allocation.

Conclusion

NPHI Database System Discussion

Public Health Implications

Health systems rely on the people within in them. No matter what technologies are available, the lack of a skilled workforce will hinder a health system's ability to detect, prevent, and properly respond to issues. With an effective training program, an NPHI can play a significant role in improving health outcomes for entire populations.

Implementing a database system will be highly impactful in this regard. Because the database can link different sources of information, business intelligence will improve when it comes to its training programs. Linking educational background with performance during the training can improve recruiting. Linking coursework and enrollment with post-training career outcomes can improve training. Information sources that have traditionally been siloed can be combined in a database to unveil invaluable insights.

All of these benefits that a database would offer will undoubtedly improve the NPHI's training efforts, resulting in a public health workforce better equipped to address health challenges of the populations they serve.

Next Steps and Considerations for Implementation

1. The first step to be taken in this project is to continue working with NPHIs to make sure that all stakeholders are able to provide input.

2. One suggestion that should be considered is a standard training database system to facilitate collaboration between NPHIs. This can be done at a regional level. For example, the Africa Centres for Disease Control and Prevention serves as a governing body for all NPHIs on the continent. Perhaps, in the future a training database could be housed here so that all training information is stored centrally. This will be beneficial as many African countries share resources, trainees, and trainers.

FluSurv Discussion

Public Health Implications

Thorough surveillance is needed to identify outbreaks as soon as possible. And outbreak investigation prevents outbreaks from becoming epidemics and epidemics from becoming pandemics. Together they are the cornerstones of epidemiology and public health. FluSurv aims to not only support both but to support both together within the system, potentially reducing the time surveillance data is received and facilitating the work of those investigating outbreaks.

Next Steps and Considerations for Implementation

1. Development and software operation (DevOps) is a relatively new software engineering practice with new to tools coming out to support it every day. More research into existing DevOps tools and how they can support FluSurv needs to be done.
2. The project focused on an emerging Type A influenza. However, FluSurv could potentially be used for any infectious disease. Looking into what steps are universal

for outbreak investigation and what steps vary based on the infection would be necessary in expanding the application of FluSurv beyond this particular influenza.

Works Cited

Fond, M., Volmert, A., & Kendall-Taylor, N. (2015). Making Public Health Informatics Visible.

Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. (2013). *The "big data" revolution in healthcare. Accelerating value and innovation*.

Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science, 332*(6025), 60-65. doi:10.1126/science.1200970

International Association of National Public Health Institutes. (2017). Who We Are.

MicroStrategy. (2017). Overview of a logical data model. *The Logical Data Model*

OpenLearn. (2016). The Database Development Life Cycle.

Whittaker, M., Hodge, N., Mares, R. E., & Rodney, A. (2015). Preparing for the data revolution: identifying minimum health information competencies among the health workforce. *Human Resources for Health, 13*, 17. doi:10.1186/s12960-015-0002-x