Massachusetts Department of Public Health

March 2015

Feasibility Proposal and Implementation Plan for a Public Health Data Warehouse

Monica Bharel, MD, MPH
Commissioner, Massachusetts Department of Public Health

Purpose of the Report

The following report is hereby issued pursuant to Chapter 284 of the Acts of 2014, Section 102 as follows:

SECTION 102. The department of public health, in consultation with the center for health information and analysis, shall develop an implementation proposal and spending plan to create a data warehouse linking relevant private and public data systems in order to receive near real-time data feeds from vital records, hospitals and other clinical partners. In the proposal, all efforts shall be made by the department and the center to limit duplicative reporting requirements by vital records, hospitals and other clinical partners. The proposal shall: (i) streamline the operation of applicable institutional review boards; (ii) engage academic partners to help support surveillance and evaluation activities; (iii) amend the department's reporting functions in order to allow for expedited reporting based on partially complete but statistically reliable data; and (iv) set forth the timeline for implementing the data warehouse. The warehouse shall be subject to the federal Health Insurance Portability and Accountability Act of 1996, 42 CFR Part 2 and all other applicable state and federal laws governing the confidentiality of personal data.

The department, in consultation with the center, shall submit the implementation proposal and spending plan, as well as any additional legislative language necessary to implement the data warehouse project, not later than December 1, 2014, to the house and senate committees on ways and means, the joint committee on public safety and homeland security, the joint committee on health care financing and the joint committee on public health.

March 13, 2015

Representative Brian S. Dempsey, Chair, House Committee on Ways and Means
Senator Karen E. Spilka, Chair, Senate Committee on Ways and Means

Representative Harold P. Naughton, Jr., Chair, Joint Committee on Public Safety and Homeland Security
Senator James E. Timilty, Chair, Joint Committee on Public Safety and Homeland Security

Representative Jeffrey Sánchez, Chair, Joint Committee on Health Care Financing
Senator James T. Welch, Chair, Joint Committee on Health Care Financing

Representative Kate Hogan, Chair, Joint Committee on Public Health
Senator Jason M. Lewis, Chair, Joint Committee on Public Health

State House
Boston, MA 02133

Dear Honorable Chairs,

Pursuant to Chapter 284 of the Acts of 2014, Section 102, the Massachusetts Department of Public Health (DPH), in consultation with the Center for Health Information and Analysis (CHIA), presents the attached ***Feasibility Proposal and Implementation Plan for a Public Health Data Warehouse*** for your review and consideration.

The Public Health Data Warehouse (PHDW) will enable researchers to examine public and private data systems from multiple sources in context with one another more quickly than ever before. Information from vital records (e.g. birth and death records), insurance claims data, public health programs, hospitals, and other clinical partners can be aligned and examined to answer critical questions about health outcomes, program effectiveness and health care costs. This effort will *substantially* advance the capacity of the Department to support policy and decision making and allocate funds more efficiently. Data reports available through the PHDW will be available to policy makers, public health stakeholders, and the public so we, as a Commonwealth, can make evidence-based and data-driven decisions that improve the health of our constituents and our communities.

The creation of the PHDW will establish a secure access point for data, while protecting and in fact, enhancing privacy and confidentiality, a critical next step for agencies to take in the face of heightened security threats. It will enable the Department to work more closely in collaboration with state agencies, including the Center for Health Information and Analysis (CHIA) to evaluate programs and examine health care outcomes in context with health care costs. This enhanced collaboration, along with our work with academic partners and other organizations acting in the public interest, will help us design better and more effective programs to improve population health. In this effort, DPH is guided by social and legal

contracts that exist between government and the public to assure the privacy, confidentiality, and appropriate management of individual health data required to protect and improve the health of the whole population.

While there are national models and guidance for this "data warehouse" concept, Massachusetts would be one of the first in the nation to develop a warehouse with this level of data and operability with other data sources that may live outside of DPH and the state system. In creating this report, DPH has worked to provide the most accurate proposal possible. It should be noted, however, that creation of the PHDW will require additional examination of options and thus may require further updates to this plan proposal. It's important, therefore, to consider some assessments presented in this document as preliminary or requiring input from a broader array of stakeholders than was possible in the available time. In particular, the estimated budget will need refinement as the Data Warehouse model is defined. Also, issues around privacy and confidentiality must have greater input from data security experts and the general public.

The following report proposes a four-year budget between $13.5 and $17.5 million to build a Public Health Data Warehouse. This proposal outlines the technical specifications of the PHDW and broad governance principles that will guide our work and the work of partners.

DPH is excited by this unique opportunity to enhance our capability to conduct the essential functions of public health and improve health care delivery and outcomes in the most cost effective ways. It will enable us to work with clinicians, insurers, and other stakeholders to improve the health of all the residents of the Commonwealth.

Sincerely,

Monica Bharel, MD, MPH
Commissioner, Massachusetts Department of Public Health

Public health data in Massachusetts is scattered in dozens of locations. To adequately plan and evaluate public health programs and policies, a data warehouse must be developed that brings these disconnected pieces together.

Background, Need, and Current Status

The Massachusetts Department of Public Health's (DPH) mission is "to prevent illness, injury, and premature death, to ensure access to high quality public health and health care services, and to promote wellness and health equity for all people in the Commonwealth." To accomplish this, DPH develops and administers programs to address specific diseases and conditions, promote wellness, and address the needs of vulnerable populations. DPH also develops, implements, promotes, and enforces policies to assure that the conditions under which people live are most conducive to good health, enabling them to make healthy choices for themselves and their families.

"One of the most important things DPH must do is enhance its ability to serve as a central repository for data and translate data for external stakeholders to set goals and track improvements."

David Seltz, Executive Director
Health Policy Commission

In order to achieve this mission, DPH must have timely and accurate data on programs, policies, and populations. While major insurers, large medical groups, and other government agencies are already developing their own data warehouses, public health data in Massachusetts is scattered across 300+ largely disconnected data systems.

This report is proposing a timeline and the financial resources needed to build a Public Health Data Warehouse (PHDW) that pulls together these various data systems and, ultimately, works with other clinical, insurance, and community data warehouses to answer critical public health questions. Legislators, researchers, public health planners – and all people living and working in the Commonwealth – will be able to get timely answers to critical questions such as: Do increases in health expenditures result in health improvements? Are Massachusetts residents healthier? What are the issues that most impact people's health? Are there disparities? If so, why and where? Are the programs funded by DPH effective in terms of return on investment and the health outcomes they deliver?

The questions facing public health analysts and researchers are more complex than ever before. DPH is no longer asked to simply track the incidence of disease. Increasingly, grants and legislation call upon the Department to assess the effectiveness of programs and policies. Even with the retraining of staff and the refocusing of programs, it is unlikely that DPH will have sufficient staff to undertake the complex analyses now required. The Department's ability to engage academic partners that support our surveillance and evaluation activities will be crucial.

Although DPH has the expertise to tackle these important questions, technology has advanced at a rapid pace, leaving the analytic tools available to the Department behind the curve. For

example, the implementation of health reform at the state and federal level has led to an enormous increase in the development, use, and transmission of electronic health data. To facilitate this change, new cloud computing technologies have been developed and implemented. Unfortunately, data breaches have also become more commonplace as hackers have learned to bypass modern security systems.

To lay the groundwork of this effort (and in response to Chapter 284 of the Acts of 2014, Section 102), the **Department has reached out to medical schools and schools of public health across the Commonwealth** to map out strategies for collaborative work using a Public Health Data Warehouse. With the PHDW, approved researchers at DPH and elsewhere will be able to query information from major datasets managed by DPH (e.g., birth, deaths, and telephone surveys on health) along with datasets managed by CHIA (e.g., the All Payer Claims Database and data from hospitalizations and emergency department visits). In some cases, clinical information from electronic health records presented in aggregate from participating health care providers may also be queried to consider health outcomes alongside the costs for providing that care.

A small-scale version of the PHDW, with limited data sets, is already enabling analysts at DPH to evaluate some aspects of the Prevention and Wellness Trust Fund (PWTF). This innovative program focuses on connecting health care and community services in nine geographic areas that encompass 15% of the population of the state. The ability to review clinical data in aggregate along with referrals to community services, health outcomes, and costs, Department and academic researchers can determine if the PWTF has achieved its goals of improving health outcomes while reducing costs and health disparities. This evaluation and subsequent adjustment of local programs is a key component of the PWTF. It will enable DPH to determine if investments in prevention and community-based care delivery will save money and actually make people healthier. The design of the PHDW will be informed by this PWTF work and its results.

> DPH has outstanding academic partners including Harvard, Tufts, Boston University, Boston College, UMass and Brandeis – which keeps them on the cutting edge."
>
> *Charles Deutsch, Sc.D.*
> *Harvard Catalyst, Harvard Medical School*

Currently, it is difficult for DPH researchers to conduct such complex evaluations. The Department's 300+ different internal data sources have been developed by individual programs using a variety of different formats. They are managed by these different programs and reside on different servers. In many cases, researchers are unable to access and examine information in different datasets to obtain the fullest picture of the effects of a program (for instance, treatment received in a clinical setting and corresponding costs). The PHDW would enable DPH and approved researchers to examine different data sets in context with one another to answer these questions.

In response to directives contained in Chapter 284 of the Acts of 2014, Section 102, **DPH has already begun to streamline the approval process for researchers**. All aspects of the current application process are undergoing a thorough examination. The process will be clarified and simplified while maintaining strict guidelines for ensuring data security, confidentiality, and appropriate use of public health data. The Department's recently released Strategic Plan has set a goal to reduce the approval time of applications for use of DPH data.

DPH is already developing new strategies to allow for expedited reporting of critical data. Algorithms developed by DPH researchers utilize partially complete, but statistically reliable, data to answer time-sensitive requests about infectious disease, opioid abuse and other topics. This process, which will be enhanced in the PHDW, will support timely public policy and fiscal decision making.

Given the vast amount of data to be queried through the PHDW, a strong privacy and data security plan is essential. Improving and updating overall data system security and capacity is a critical challenge for large data-rich agencies, like DPH, as data security concerns become increasingly more important. With the creation of the Data Warehouse, these systems will be revised, streamlined and updated.

Access to and operation of the PHDW will be governed by all applicable state, federal, and agency privacy and security regulations. Researchers will *never* be allowed to see individual records through the PHDW. Even aggregate data about groups will be strictly controlled. The PHDW will be administered by the Commissioner of the Department of Public Health, who will be advised by a group of independent experts, ensuring needed oversight, ethical and legal clarity and fiscal accountability.

Development of this PHDW report has involved thoughtful identification and consideration of many different issues. These include:

- Technical specifications to ensure functionality, data security, and the efficient use of the state's resources;
- Plans to protect the privacy and confidentiality of individuals whose data is included;
- Improving processes for users to obtain secure, timely access to data for projects in the public interest;
- An ethical framework for linking data and conducting analysis;
- Resources necessary for implementation and future operation of the warehouse.

By improving access to high quality data on health outcomes and health care costs, DPH looks forward to this opportunity to help policymakers, insurers and other stakeholders to adjust plans, programs and budgets to achieve better health and cost outcomes that improve the health of all Massachusetts residents.

Technical Specifications

Goal: *To design a technical architecture for a data warehouse that utilizes state resources efficiently while enabling secure access to public health data for internal and external users.*

Background: DPH maintains and/or has access to more than 300 separate data systems that span more than 100 programs, each with unique but important missions, goals and data requirements. Because of the differences among these datasets, DPH has very little ability to combine queries of more than one dataset or to query these systems for purposes outside of their original design. Without that ability, answering critical and timely public health questions is severely limited. The Data Warehouse will provide standardized formatting for copies of these datasets so that researchers can quickly access multiple data sets to answer critical public health questions. It will deliver answers to questions such as: What are the impacts of policies and programs on the health of people living in the Commonwealth? Have Department programs resulted in health care cost savings? Which programs are most effective in terms of health outcomes? These questions cannot be fully answered without a data warehouse.

Several close partners of DPH are already working on data warehouses that allow them to conduct similar sophisticated research.

- The Center for Health Information and Analysis (CHIA) is currently building a data warehouse which will contain two datasets of significant interest to the Department: the All Payer Claims Database (APCD) and the Acute Hospital Case Mix Databases (CASEMIX). The APCD contains the medical claims of virtually every insured resident of the state since 2009. CASEMIX has basic hospital event data from all acute care hospitals for over a decade. CHIA currently queries these datasets to analyze health care costs.

- Harvard Vanguard/Atrius (HVMA) maintains a data warehouse of clinical data, as well as a limited number of medical insurance claims. HVMA has partnered with the Department to use this data to analyze the impact of tobacco intervention systems on the health of its patient population.[1]

- The Massachusetts League of Community Health Centers (MLCHC) operates a data warehouse of clinical information that allows community health centers to query their data for quality improvement purposes.[2] The League has partnered with DPH to support work evaluating the impact of programs such as the Community Transformation Grants, Mass in Motion, and the Prevention and Wellness Trust.

---

[1] Land TG, Rigotti NA, Levy DE, et al. (2012) The Effect of Systematic Clinical Interventions with Cigarette Smokers on Quit Status and the Rates of Smoking-Related Primary Care Office Visits. PLoS ONE 7(7): e41649. doi:10.1371/journal.pone.0041649.

[2] For further information, see http://www.massleague.org/About/FactsIssuesBrief.pdf

- The Massachusetts League of Community Health Centers, Harvard Vanguard Medical Associates and Cambridge Health Alliance have coordinated with DPH and Harvard Medical School's Department of Population Medicine (DPM) to create a data warehouse of clinical data called MDPHnet.[3] MDPHnet uses a federated structure, which allows organizations to maintain possession of their own data while permitting external access to approved researchers. When DPH analysts query primary care clinical records of 1.2 million patients, the results are aggregated and reported back to analysts without DPH having to take possession of – and store and secure - this protected health information.

The Department of Public Health has a wealth of data but no Data Warehouse. To be meaningful, research on public health questions must simultaneously examine the effects of public policies, community infrastructure, community services, insurance coverage and claims, clinical measures, and personal behavior. Current projects at DPH that utilize data from different systems must be planned and developed from scratch. Often, code cannot be shared because data formats are incompatible. Few systems talk to each other.

Recommended Plan: *DPH should build a data warehouse to manage its own data. It should implement a plan that connects the DPH warehouse to other relevant Warehouses such as the one being developed by CHIA. This federated or virtual data warehouse will maximize the use of existing infrastructure among key partners while supporting critical public health evaluation needs. Each participating organization would maintain possession and control of its own data, while allowing access to it by approved researchers.*

There are four primary reasons to support this recommendation:

First, a federated or virtual model would be cost effective, eliminating the need to copy and maintain two separate data warehouses to contain the vast APCD data files housed at CHIA (>10 terabytes). CHIA has purchased the necessary hardware to hold these very large files and is already working with DPH to establish the legal framework to provide DPH access to APCD and CASEMIX for approved projects. A federated DPH data warehouse will not duplicate the work of CHIA, but rather will enable the Department to work more efficiently and more closely with them, accessing data already in existence.

Second, the use of existing infrastructure will speed the completion of the PHDW development work. Using existing state systems, we can eliminate record duplication and match records despite subtle variations (for example, field names titled "address" and "street address" that can represent the same information.)

---

[3] Klompas,M; McVetta, J; Lazarus, R; et. al. Integrating Clinical Practice and Public Health Surveillance Using Electronic Medical Record Systems. American Journal of Public Health: June 2012, Vol. 102, No. S3, pp. S325-S332. doi: 10.2105/AJPH.2012.300811.

Third, clinical providers across Massachusetts and the United States have adopted Electronic Health Record (EHR) software systems in record numbers,[4] primarily due to incentive payments offered by the Center for Medicare and Medicaid Services (CMS) as part of the Affordable Care Act of 2009 (ACA). These records contain valuable information on diagnoses, treatment, medications and other variables. A federated data warehouse will allow clinical partners that choose to work with DPH to maintain control of their electronic medical records at the same time that they allow data queries from DPH and participate in work designed to improve the health of their patient populations. DPH has already partnered with providers across the state to support public health evaluation and quality improvement projects by utilizing clinical data.

Fourth, federated models have been thoroughly tested and are in wide use. They provide secure and reliable access for sensitive information for a wide variety of organizations across the US and around the world.

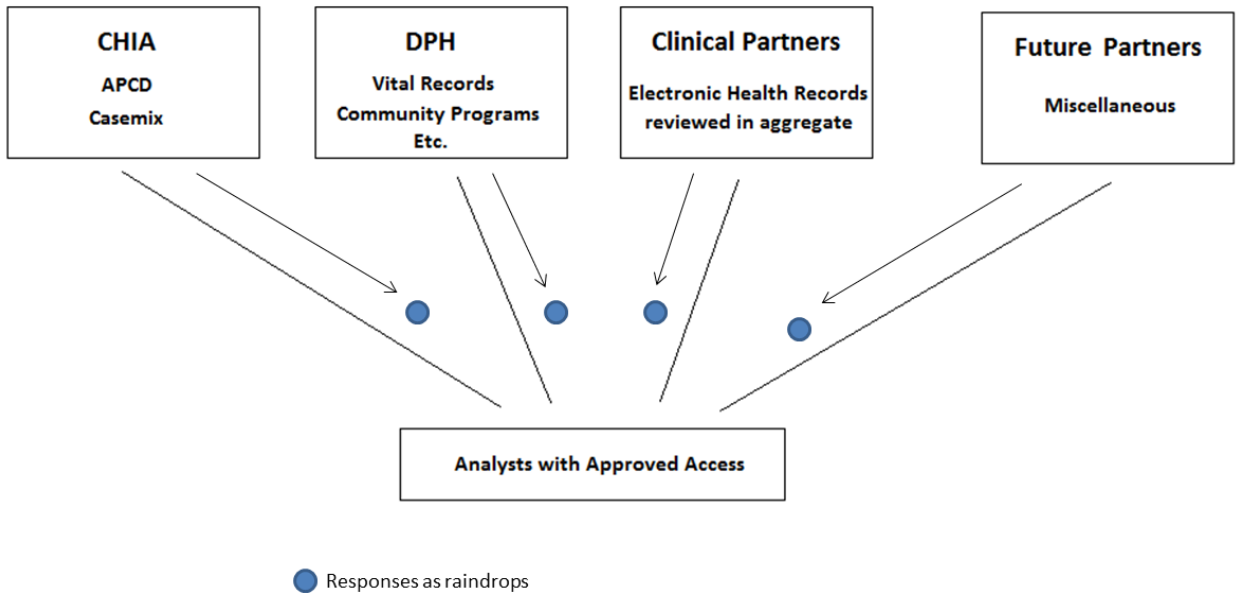
Case Study #1: A major premise of the Prevention and Wellness Trust (PWTF) model is that extending care into the community can yield measureable health improvements and cost savings. Falls among older adults is one of the priority health conditions for the PWTF and local grantees are already providing services related to reducing the risk of falling. Currently, no data system within or outside of DPH captures the breadth of information required to determine the effectiveness of the interventions funded through PWTF. To do so would require linking clinical screening data for falls risk, community services provided, costs associated with injuries from falls, and deaths from falls. Clear demonstration that extending care into community settings has impacted clinical and cost outcomes is critical for Massachusetts and the nation. A successful outcome means that the PWTF model can be a tool for controlling health care costs. Without the DPH Data Warehouse, an optimal demonstration of the utility of this approach is not possible.

How it Works

Data resides in more than one location, remaining with its "owner" agency. The analysis of data, therefore, would rely on cloud-based technology or cloud computing. In a sense, each data system would reside in its own cloud or online location. In the proposed DPH model, analysts whose projects have been approved would submit code to answer questions related to their projects. Relevant data from these separate clouds would come together briefly, be evaluated using sophisticated statistical models, and then presented in summary form in order to answer the analyst's original question. Where feasible, we will utilize the Mass HIway (the Massachusetts Health Information Highway) to transport information through the DPH Data Warehouse.

[4] For information, see http://dashboard.healthit.gov/quickstats/

To expand the cloud analogy, this process can be thought of as "raindrops" of data that gather to answer questions, and then evaporate as soon as the answers are delivered back to the analyst. Data raindrops hold together for as brief a time as possible, thus limiting the risk of release of private health information. The figure below depicts the proposed model.



As part of developing the infrastructure to evaluate the Prevention and Wellness Trust Fund, DPH has already developed a small-scale data warehouse called a "staging area." Currently, analysts can pose very basic questions, test data comparison models, standardize formats, and otherwise prepare data for delivery to a larger data warehouse. The staging area contains a reduced-size set of the APCD and Massachusetts death records from the Registry of Vital Records and Statistics (RVRS). DPH epidemiologists are examining algorithms that will enable smooth connection of DPH data systems to CHIA and are already querying these datasets to test compatibility and connectivity.

Case Study #2: Examining or projecting the impact of environmental changes or new public works projects is also possible through the PHDW. For example, we could look at the health and cost implications associated with the development of a regional airport. The PHDW would permit us to study the effects of an airport after it was built or even estimate the impact of a proposed development. Currently, it is difficult for analysts to tie health effects and increases in costs to any environmental change. To understand whether a new airport has impacted health and healthcare costs, it would be necessary to simultaneously look at changing pollution and noise levels, ED visits, hospitalizations, medication use, and specific healthcare costs. Clinical measures might focus on hypertension. Furthermore, to relate any change in clinical or cost outcomes, comparison populations would be carefully matched to the population most affected by the new airport. Without a proposed PHDW, these analyses could not be done and thus important health and policy conclusions could not be drawn. A PHDW would enable simultaneous access to relevant data and permit researchers to examine this complex problem using appropriate tools.

Key Take-Aways
(Technical Specifications Section)

- Many partners already have data warehouses
- DPH data systems are largely disconnected
- Cloud computing and federated data access is proven technology
- Utilizing existing infrastructure will make the PHDW cost efficient

Summary: A federated data warehouse is both efficient and cost effective. Utilizing existing infrastructure will ensure that people living within the Commonwealth realize the benefit of the PHDW as quickly as possible. Clinical partners are in a better position to share electronic data than ever before. Some sharing of clinical data is already taking place through the MDPHnet system, but this is more limited than what is anticipated through the PHDW. Finally, a federated model is based on existing technology that has been thoroughly tested and vetted.

Privacy and Confidentiality

Goal: *To minimize the likelihood of inadvertent release of private health information.*

Background: In any plan to link and utilize large amounts of health data, maintaining privacy and confidentiality must be of primary concern. Data breaches that tend to receive the most attention are those that involve personal financial information, credit card numbers, or social security numbers. Although significant attention has been paid to breaches at Target, Home Depot, and other retailers, the health records of nearly 30 million patients in the United States have also been breached since 2009.[5] In the design and operation of the data warehouse, DPH's goal is to minimize the risk of inadvertent release of private health information.

Intentional and accidental breaches of data are not the only pathways to release private health information. Recent studies provide guidance regarding the best options for and limitations to protecting personal data.[6,7] Given the sensitivity of the data accessible through the proposed warehouse, several additional design requirements will be implemented to ensure private health data is maintained, the risk of inadvertent release of information is minimized, and private data remains private.

Recommended Plan: DPH will follow eight procedures to ensure that privacy concerns are adequately addressed.

1. To ensure that private data remains private, DPH proposes to follow federal NIST and HIPAA security rules. These may also include use of *state of the art 256 bit encryption* for all data in the PHDW, as well as single-use keys linking records across tables. Evaluation of system performance will inform decisions about what level of encryption to employ to maximize security and functionality.

2. *Analysts will be prevented from seeing patient-specific data*. A privacy shield will prevent users from seeing patient-specific data owned by any other organization.

3. All analysts, project managers, and technicians will be required to complete an *annual data privacy and confidentiality training*. This applies to DPH staff, other government agencies, and any external partners that access data through the Data Warehouse.

4. *All direct personal identifiers (name, address, etc.) will be housed in a separate location* not accessible to any analyst or project manager whether internal or external to DPH. Names, addresses, and social security numbers will never be included in any data set

[5] For more information see http://www.washingtonpost.com/blogs/wonkblog/wp/2014/08/19/health-care-data-breaches-have-hit-30m-patients-and-counting/

[6] Curfman GD, Morrissey S, Drazen JM (2011) Prescriptions, privacy, and the First Amendment. N Engl J Med 364: 2053–2055.

[7] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

accessible through the warehouse. No credit card number, phone number, bank number, or any financial information will ever be included in any data set. Other protected health information (PHI) such as zip code or birthdate will only be stored in an encrypted format.

5. *Data queries will be restricted to approved uses only*. This type of access is generally termed role-based access. With role-based access, researchers or other applicants can query and analyze only that specific data that has been requested and approved for a particular study. DPH and CHIA have separate approval processes, which will be maintained. Instead of providing entire copies of datasets, the PHDW will provide the ability to query only specific data sets approved for each project.

6. *Cell suppression will be automatic*. The publication of reports with small numbers of cases can make it possible to identify individuals in large and small data sets. By requiring automatic cell suppression, small cell sizes will not be seen by analysts who submit queries and thus cannot be reported. Implementing automatic complementary cell suppression will eliminate the risk of inadvertent disclosure of protected information.

7. *Geo-coded data will be protected through spatial confidentiality algorithms*.[8] Geo-coding data can be a powerful analytic tool that enables visualization of effects as well as adding significant context from the communities in which people live and work. (Example: the number of a particular type of cancer cases in a specific city or town). At the same time, geo-coding can sometimes identify individuals as effectively as names, addresses, and social security numbers. Spatial confidentiality algorithms prevent this by adjusting the location of geo-codes randomly in every direction making re-identification from geo-codes alone virtually impossible. In addition, geo-codes used in individual projects will not be stored in the Data Warehouse. They will be housed with other direct identifiers in a separate location and used only to capture relevant community and neighborhood data associated with the geo-coded location.

8. *Thresholds for differential privacy will be set.* Differential privacy is a term describing the method for calculating the risk to any individual's privacy when s/he is included in a database.[9] The notion of differential privacy is that thresholds can be set to raise or lower the likelihood that any individual can be identified from reported data. DPH will explore and set appropriate thresholds to provide assurance that private data remains private.

---

[8] Curtis AJ, Mills JW, Leitner M. Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina. Int J of Hlth Geographics 2006, 5:44 doi:10.1186/1476-072X-5-44.

[9] Dwork, C. http://www.msr-waypoint.com/pubs/64346/dwork.pdf

Key Take-Aways
(Privacy and Confidentiality Section)

- Maintaining confidentiality and privacy are primary concerns
- DPH will follow federal NIST and HIPAA security rules
- Analysts will be prevented from seeing patient-specific data
- Responses to queries will only include aggregate or summarized data
- Only approved users will be able to access data

Summary: While the PHDW will provide strict security to prevent improper access, DPH will also take additional steps to ensure that privacy and data confidentiality are preserved. Many of these measures overlap; they provide multiple layers of protection for private health information. Foremost among these measures is the fact that analysts will not be permitted to see data. They cannot report (and thus disclose) what they cannot see. Analysts will only be able to access data that has been approved for a specific project. Results with small numbers will be automatically suppressed. Privacy training will be required annually for all users. Personal identifiers including geo-codes will be housed in a location separate from the warehouse. Sophisticated algorithms will be used to set thresholds to further ensure that the likelihood of an inadvertent disclosure of data is minimized.

Data Warehouse Governance Principles and Execution

Goal: *To develop a governance structure for the DPH Data Warehouse that provides relatively straightforward access for internal and external users at the same time that privacy, ethical and legal issues are addressed.*

Background: From the beginning of this effort, building the ethical frameworks and processes necessary to support the DPH Data Warehouse has been a key concern. Our work started with recognizing that DPH is guided by social and legal contracts that exist between government and the public to assure the privacy, confidentiality, and appropriate management of individual health data required to protect and improve the health of the whole population. We also recognized the important ethical imperative associated with making use of emerging technologies and analytic opportunities.

In preparing this report, DPH investigated new ethical, legal, technical, and policy resources that address the acquisition, storage, and appropriate use of publicly and privately held health information.

Recommended Plan: DPH will develop a governance structure that embodies a set of core assumptions for the ethical management of access and use of private health information. DPH, CHIA, academic, and other community partners shall participate in an advisory group to ensure that the public's interest is maintained.

Core Assumptions for Governance of the Data Warehouse

The development of this PHDW has been guided by a set of Core Assumptions to ensure that all policies, procedures, and proposed projects are ethically sound and grounded in the public interest:

- There is a communal benefit from and responsibility for building a learning health system that can better serve the health and public health improvement needs of the Commonwealth.

- Increased data linkage and analytic capability can - and must - add value to needed population health improvement knowledge and strategies.

- The integrity of the data infrastructure, linkages, access, and analytic mechanisms is the platform upon which the appropriateness of this enterprise sits.

- Transparency in purpose, process, outcomes, evaluation and monitoring must be reflected in easily accessible and understandable information.

- The governance structure must act in service to all these criteria.

Structure

The DPH Office of Data Management and Outcomes Assessment (DMOA) assembled a small group of experts to conduct a review of data warehouses already established by other states, federal agencies, academic institutions, and clinical organizations. Particular focus was paid to the following:

- The intended purpose of each data warehouse
- How each data warehouse is governed
- Datasets included in each warehouse and how they are linked
- Control and access to the data
- Vetting and approval processes
- Protecting the privacy of individuals whose records are contained in the data warehouse
- Types of organizations contributing and accessing data
- How each data warehouse is funded

Governance of the Massachusetts project must closely reflect the state's commitment to scientific rigor, data security, and protecting the privacy for individuals whose data is included. DPH will create an advisory group as part of the governance structure for the PHDW. This group will be composed of state and national experts and representatives who will:

- Advise on operations and scientific activities of the Data Warehouse
- Work with the Department to define ethical and legal policies and procedures for data access and use
- Ensure that proposed projects serve the public interest
- Act as a forum for communicating the activities of the PHDW to external partners
- Foster collaboration among state agencies, academic institutions, and other organizations working in the public interest to optimize surveillance, program evaluation, and research throughout the Commonwealth.

Key Take-Aways
(Governance and Ethics Section)

- Use of PHDW will be guided by core assumptions. Among them are:
  - Use of public health data should serve the health needs of the Commonwealth
  - The purpose and monitoring of the PHDW must be transparent and easily understood by the public

- DPH will create an advisory group of state and national experts and representatives to guide work with the PHDW. They will:
  - Advise on operations
  - Define ethical and legal policies
  - Ensure public's interest is maintained
  - Communicate with external partners

The advisory group will consist of three standing committees:

- The **Scientific Committee** will advise on methods for designing studies and analyzing data, as well as identifying potential additional resources and collaborative projects.

- The **Operations Committee** will advise on developing appropriate application forms, policies and procedures for overall operations of the data warehouse, including data security.

- The **Ethics and Public Interest Committee** will advise on developing and implementing appropriate policies and procedures to ensure the data warehouse embeds protection of privacy and thoughtful consideration of ethical and legal issues. Consumer representative(s) will be included among the members of the Ethics and Public Interest committee.

Estimated Budget and Timeline

Goal: *To develop a cost effective and sustainable plan for adding data systems to the DPH Data Warehouse that enables analysts and researchers to address critical public health evaluation questions as quickly as possible.*

Recommended Plan: The DPH Office of Data Management and Outcomes Assessment (DMOA) has determined the core public health data sets that would be considered for inclusion in a Public Health Data Warehouse. This Office intends to include nearly all of DPH's core data in the PHDW at the end of four years of funding.

Please note that the budget estimates presented below are preliminary. To arrive at these estimates, DPH used standard rate and time estimates for personnel required to complete each phase of this work. Data sets were ordered by level of complexity with respect to the information included. DMOA also made an initial assessment of the complexity of linking individual data sets to other data sets and included this information in estimates of the time required to complete this work. Given that the Department's knowledge about this work will increase over time, estimates will also be refined as more information becomes available and the "bottom line" may change.

Building a PHDW involves three high-level IT phases that are implemented simultaneously. Phase 1 is building the Operational Data Store. That involves making the data ready for accessibility through the data warehouse, but not linking it to any other dataset. This work will continue until nearly all of DPH's core data are accessible through the PHDW. Phase 2 is building an Enterprise Data Warehouse using a federated model. This phase consists of aligning variables within the datasets to allow linkage of variables. The third and final phase is the Presentation Phase, which builds in suppression parameters to meet privacy and other confidentiality requirements.

These phases would be implemented simultaneously, with an initial focus on building the infrastructure and completing all three phases for initial core datasets. Each following year, additional datasets would be added to the core.

Despite the fact that the Department has proposed a four-year timeline to complete the Warehouse, DPH analysts and other researchers will be able to query data as soon as it is incorporated in the PHDW. In other words, important public health questions can be addressed and answered throughout the construction phases.

Below is an estimate for the overall production and operations costs for building the Public Health Data Warehouse.

- Averaging standard Commonwealth of MA, Operation Service Division IT Contractor rates for positions and multiplying the projected project hours by the hourly rates, the total

estimated cost for the estimated 150,000 hours of labor is in the **$12,000,000 to $14,000,000** range.

- Industry standard estimates for the data center costs for multiple environments that support Development, Test, Quality Assurance (QA) and Production (internal and external facing environments) for the hardware (including networking, security, web, application and database servers, storage, back up and disaster recovery), software (robust warehousing, reporting and middleware) and the operational support staff to manage and maintain the environment would range from **$1,500,000 – $3,500,000** or the course of a 4 year project.

- The total IT cost not including office space, IT equipment for office use (i.e., laptops, desktops, mobile devices), and 'in kind' cost to existing EHS IT employees is estimated to be in the range of **$13,500,000 to 17,500,000**.

- The above cost does not include the cost for DPH Executive, Program, Operations or Legal and Oversight Staff.

Summary: The total estimated budget over four years ranges from $13.5 to $17.5 million. Implementing the three phases simultaneously includes building the infrastructure and making core data available in the PHDW. The costs to achieve this include personnel, hardware and software, and hosting and maintenance costs. The first year emphasis would be on building the infrastructure and completing all 3 phases for initial core datasets. Each following year, additional datasets would be added to the core.

Key Take-Aways
(Budget and Timeline)

- Three phases of work will take place simultaneously:
  - Operational Data Store
  - Enterprise Data Warehouse
  - Presentation Phase
- The 1st year will emphasize building the infrastructure and initial core datasets
- Work will continue until nearly all of DPH's core data are accessible through the PHDW
- Researchers will be able to query PHDW throughout construction phases
- Total IT costs over 4 years for the PHDW estimated to range from $13,500,000 to $17,500,000
- A Public Health Data Warehouse Trust Fund should be established

Public Health Data Warehouse Trust Fund:
"There shall be established upon the books of the commonwealth a separate fund to be known as the Public Health Data Warehouse Trust Fund to be expended, without further appropriation, by the department of public health. The commissioner of public health shall, as trustee, administer the fund. The fund shall consist of revenues collected by the commonwealth including: (i) any revenue from appropriations or other monies authorized by the general court and specifically designated to be credited to the fund; (ii) any funds from public and private sources, including gifts, grants and donations; (iii) any interest earned on such revenues; and (iv) any funds provided from other sources. The department may incur expenses and the comptroller may certify for payment amounts in anticipation of expected receipts, but no expenditure shall be made

from the fund that would cause the fund to be in deficit at the close of a fiscal year. Monies deposited in the fund that are unexpended at the end of the fiscal year shall not revert to the General Fund."

Summary

As is often the case, Massachusetts is leading the way in innovative health policies. In this era of health care reform, the nation is looking to the Commonwealth to guide the discussion of which policies, programs, and approaches can improve the health of the population and reduce healthcare costs. The answers to these critically important questions are as complex as any that public health has ever encountered.

In order to adapt and thrive in this new era, DPH must upgrade its data management system and take advantage of new technologies to face the public health and financial challenges of the 21st century. New and emerging technologies offer rich opportunities to collect, access, analyze and secure data.

This plan details an approach to incorporate key data systems, with strict security, legal and ethical measures, into a DPH federated Public Health Data Warehouse. This will enable researchers to query and analyze data, evaluate policies and programs, and provide timely and accurate information to legislators, policymakers, funding sources and the general public. DPH will work with researchers and analysts inside and outside of state government to develop the quality measures and cost evaluation that will lead to improved health for all people who live and work in the Commonwealth.

The plan presented here capitalizes on tested methodologies such as federated warehouse models and confidentiality procedures. Nonetheless, there is also a recognition that the work in the area of cloud-based computing will continue to evolve. Any final plan for the PHDW will involve additional examination of options which could necessitate revisions to this plan.

The estimated cost of developing and maintaining the DPH Public Health Data Warehouse for 4 years is $13,500,000 to $17,500,000.