Administrative Data and Disease Surveillance: An Integration Toolkit

Introduction

Each day new demands arise for more health information, whether those demands are from the public or private sector. Public health should make full use of all existing data resources to meet these new demands before looking to new data collections, given limited resources. Integration of clinical and administrative data could become an essential resource to address new questions and surveillance demands. This paper provides a framework for the process of integration—a framework that will assist public health practitioners acquire administrative data for linkage to their surveillance data systems.

Data integration is dependent on the resolution of human and technological factors—this paper focuses primarily on the human or cultural factors that must be addressed for integration to occur. Privacy issues, lack of confidence in government, turf protection, unique history of data—are part of the culture of the data. Examples from several states are included that reflect on these key issues. While the examples are useful for understanding the issues associated with integration, it is clear that each state, region, or local public health environment differs. The framework can assist others as a starting point for the necessary dialogue that must occur before data integration is possible. We do not have all the answers— it is our hope that others will enter the discussion and will suggest improvements to the framework.

The National Association of Health Data Organizations (NAHDO), through CDC funding, has undertaken this effort to promote the use of administrative data in conjunction with the existing clinical data resources, thereby assuring that administrative data are "re-used" and providing a guide to integration efforts in state and local public health.

Administrative Data/Surveillance Data: Similarities and Differences?

Many states collect electronic discharge data from hospital billing records and make these data available for public health use, research and for use by other constituents. Generally, they contain primary and secondary diagnoses, procedure codes, provider names, admission and discharge dates, and demographic information on the individual. Some states have unique patient identifiers, while others may only have identifiers unique to the care provided by that one provider. Across states, there are some differences in data elements and formats, but generally most states follow either the UB-92 or now the 837 professional claim formats. Data editing programs are used to improve the quality of the data that is submitted. The databases cover the

entire population of individuals discharged from acute care hospitals, some may also include specialty hospital discharges, or sub-acute discharges. Administrative data collection is an efficient method of acquiring healthcare related information, however, it does not have the detail available in surveillance data collections.

Surveillance databases collected by public health generally focus on a specific type of condition, such as cancer, sexually transmitted diseases, immunization, diabetes, etc. The data elements and formats differ between types of conditions, but generally there is a significant amount of detail in the data regarding the condition and the individuals' demographics. The surveillance data elements are abstracted from the medical record by healthcare providers and/or registry staff, and then transferred to an electronic format. The surveillance data contain direct identifiers of the individual, often including name and address. The data are generally sent to the CDC for national surveillance and reporting activities. This type of data collection is more expensive than the collection of administrative data.

While the surveillance databases have more detail, administrative data can add value to those systems—assisting in answering questions related to access to care, evaluation of prevention efforts, policy analysis, workforce distribution, etc. It is also possible to use data mining strategies to identify cases missed in a disease register, identify new disease patterns, study the occurrence of relatively rare conditions, and to estimate local variation and subgroup patterns. Virnig and McBean (2001)[1] describe several studies that have examined the capacity of administrative data to identify an incident of cancer found in a cancer register, or an immunization. Depending on the study, administrative data were able (in varying degrees) to identify many, but not all cases. Administrative data can also be used to validate surveys of some self-reported conditions, such as diabetes. Hebert et al (34) developed an algorithm to compare the self-reports of diabetes found in the Medicare Current Beneficiary Survey with diagnoses of diabetes in Medicare administrative data. The integration of administrative and clinical data provides us with a less expensive alternative to the expansion of clinical systems.

Existing administrative and surveillance systems are currently being challenged to address national issues such as bio-terrorism, rapid spread of contagious disease, and other environmental threats to humans. To address these challenges we must attempt boundary spanning via integration; the country cannot afford to extend the clinical databases to cover the nation.

[1] Beth A. Virnig and Marshall McBean. "Administrative Data for Public Health Surveillance and Planning," Annual Reviews Public Health, (2001), 22:213-30.

What is integration?

The need to promote integration of data systems is precipitated by a dramatic change in the past decade in information technology as well as a growing need for timely and accurate data. When many of our public health legacy systems were designed and first implemented, the hosting computer required a large environmentally controlled room. To justify the large expense to maintain these computing facilities, many if not all an organizations information systems resided on that single computer. Many of today's desktop computers have more processing power than the early computers that required that environmentally controlled room to run. These advances in computer technology have enabled new generations of public health systems to be built and maintained in a distributed environment. Along with the advances in desk top computing has been a parallel explosion in the availability of "user friendly" software tools that further enabled the development of distributed systems.

Advances in hardware and software technology have not changed the need to share information across systems. There was a time when the cost of duplicating the information to be shared was less than the cost of integrating these distributed systems. The complexity of today's public health societal responsibilities to control the costs of our information systems is making it cost prohibitive to sustain distributed public health system that duplicate information--information that needs to be shared.

In today's information system environment, data system integration is not a luxury. To sustain our increasing need for data at an affordable cost, we must design integrated distributed systems. Advances in our technology make it impractical to revert back to using only centralized databases. Shrinking work forces and the need to share information makes supporting redundant solutions equally impractical. The only realistic solution for today's shared information needs is to develop strategies to integrate the variety of specialized public health systems.

For the purposes of this paper, we are defining integration as data sharing between administrative and clinical distributed systems. The escalating need for cost effective ways to use these distributed systems makes integration necessary. The emerging and enabling new technologies makes integration possible. The success of these integration strategies will depend on our ability to design sustainable and work force neutral systems. This will be the cornerstone for establishing a trusting relationship between data users and data suppliers. Knowing that the diseases and events that need to be monitored and tracked are oblivious to political boundaries, the key to

sharing data across those political boundaries will be the use of standards in all the system designs.

Current Stressors on Clinical and Administrative Data Systems

In this section, we discuss the environmental stresses and strains on the stewards of clinical and administrative data systems. We suggest that integration efforts will be more successful if approached with an understanding of the existing constraints and new demands placed on these systems.

Factors Affecting Public Health Clinical Data Systems

Public health has a long history of monitoring the health and well being of the public, whether on the local, state or national level. Surveillance of disease is an on-going and ever expanding responsibility for public health officials; the collection and analysis of data/information is an integral component of this disease surveillance.

New surveillance activities are contributing to an increasing workload that has strained public health officials. The most dramatic impact was from 9/11—it raised the specter of the potential for an attack on our public utilities—postal services, nuclear plants, food supply, transportation, and other public services, requiring new surveillance and emergency action. The nation's public health was threatened by terrorist actions, as was the public health system. Also occurring is the spread of new cases of West Nile Virus; it has spread to the Midwest from the East Coast, much faster than anticipated. A myriad other conditions, such as: Lyme disease, outbreaks of e.coli, and tuberculosis also produce additional threats to the population. The numbers of Hepatitis and HIV/AIDS cases continue to increase, two new cases of Bubonic Plague have appeared, and now Monkey Pox; the list of infectious disease outbreaks continues to grow, including diseases once believed eradicated.

The new public health responsibilities place enormous stress on a system already coping with a wide variety of surveillance activities including lead poisoning, chronic diseases, immunizations, sexually transmitted disease, etc. Each of the existing surveillance responsibilities requires a tracking system for identifying and monitoring new cases and new tracking systems are coming on line as new conditions emerge. These additional systems are overloading the public health system, particularly at the local level where public health officials are responsible for surveillance, prevention and intervention activities.

Many surveillance systems are still "paper-based" or, if electronic, are idiosyncratic to the specific surveillance need. The different surveillance systems do not, in most cases, work similarly, nor do they produce results that are easily integrated with other systems. The databases have different designs, formats, collection tools, differing definitions for similar data elements, and are often used in isolation of each other. Many of the existing systems address only identified cases and are not designed for case-finding action in population-based systems. Yet, these systems have historically met the needs of public health officials. Now, with expanding responsibilities and with no additional workers, public health professionals are under pressure to utilize increasingly sophisticated technology for data collection and analysis. The pressure on the system is overwhelming.

At the same time as these new public health threats are increasing workloads, new healthcare data standards are in the process of implementation, as a result of the Administrative Simplification component of the 1967 Health Insurance Portability and Accountability Act. Meeting new transaction and privacy standards is required for covered entities under this legislation. Covered entities include all healthcare providers, plans, and clearinghouses exchanging electronic billing forms. The new transaction requirements have meant that many healthcare providers (who serve as a primary source for data surveillance system input) are being forced to radically change their data processing systems. In particular, healthcare providers are altering the manner in which data is formatted and shipped to other entities.

While some activities of public health are exempt from the HIPAA rules, other programs/services are not exempt. Determining the status of the programs has been difficult, since many are partially exempt and partially covered under HIPAA. State public health officials publicly report they will follow the HIPAA standards, yet many of the data systems in public health will be difficult or expensive to change, given the age of the software and the hardware. While Y2K (year 2000) efforts revamped problems with two-digit date issues, other more difficult changes are needed to comply with both the Transactions Standards and the Privacy Standards. It is difficult for public health to reconcile differences between Federal and State privacy laws and even more difficult to determine whether state or federal law preempts[2] the other.

Adding to the problems—are legislatures that are rejecting requests for new data systems for public health even though there is a legitimate surveillance need. In some states the legislators are being lobbied to reject new systems, by individuals and advocacy organizations—they feel

[2] The American Hospital Association has produced a document to assist in analyzing State Law Preemption Under HIPAA.

threatened by the sensitivity and volume of electronic medical information that is available. From their standpoint, registry information contains potentially "threatening information" —citizens are fearful of insurance and employer blacklisting or excessive insurance premiums should the registry information become known outside of public health. Exposures of health data on the Internet have alarmed citizens and they have taken their concerns to the legislatures.

Given today's concerns with privacy, integration efforts are becoming more difficult. Sensitive information should be added only when there are specific questions to be answered by addition of that sensitive information—and then the data linkage should contain only the data elements necessary to answer the specific question(s).

In the details of HIPAA Transaction Standards, specific code sets are mandated for such things as pharmaceuticals, dental care, and medical diagnostic and procedures codes. As a result of these implementation rules, there are additional conflicts beyond laws, and these directly affect integration efforts. The HIPAA standards for diagnostic codes conflict with the new national standard for the death certificate, required by another authorizing unit of the Federal system. The death certificate standards went a step further than the HIPAA standards--requiring implementation of the International Classification of Disease-10 Revision (ICD-10) diagnostic coding standard. As a result of this difference, healthcare providers will be storing and submitting information to other covered entities using the International Classification of Disease-9th Revision-CM (ICD-9-CM) standards as required by HIPAA, but submitting information to state and local Vital Records offices using the new ICD-10 diagnostic coding schema. This requires providers and those using this data for integration to maintain a dual-system of data management and reporting or complicated system designs with necessary code translation processes, until such time as the HIPAA transaction standard catches up. This is just one of any number of conflicts among data standards, resulting from the various authorities governing healthcare provider data.

In summary, public health entities face a series of new challenges from the environment—overwork, isolated data systems, legislative veto on new systems, increasing data standardization in the healthcare provider community, and changing/conflicting regulations and standards. Following a similar discussion about constraints on administrative data, we suggest why data integration is so critical to local public health.

Factors Affecting Administrative Healthcare Data

There are significant differences in the design time and energy needed to build administrative data systems versus development of surveillance clinical data. Much of the difference in design and implementation time is related to the goal associated with the establishment of the data system—public health systems are retrospective following a crisis or epidemic or to assess program effects, while administrative data systems are built prospectively to address future concerns of payers, purchasers, policy analysts, consumers, and healthcare providers. While the future is of interest to policy makers, it is just not as compelling as fixing a problem that exists. The business case for establishing a new administrative system must be clearly articulated; discussions about the merits of the new system are time-consuming. The actual implementation is arduous as well and takes considerable time to complete.

On average, it can take 5-10 years to "bring up" a single administrative data system in a state. The birthing process for an administrative data collection includes coalition-building activities, initial advocacy for the data system legislation, the legislative process, initial implementation of the data collection, and the validation of the data. The technology of collection is generally based upon proven electronic technology given the volume of records, but systems also have to be able to accommodate to the lowest level of data submitters' system sophistication. Discharge data systems may be based on a clean abstract of the claims data or may require providers to abstract and re-code data elements to meet the state authority requirements. The local recoding system creates difficulties in linking data to other states and other public health data systems. These design challenges hinder integration of data—changing just one data element may require new statutory language and rules—creating a formidable and time-consuming barrier to integration.

The design of the state-wide discharge systems, while remarkably similar to each other, have not had established standards for content of the system, format, or for definitions of data elements. Each system was designed based on how the system could best meet the needs of the authorizing body and, these systems also reflect the process of negotiation occurring in policy making entities--conflicting needs and demands between various participants alter the design of the data system even when the initial intent is to have a standard data system. For example, recently approved data systems may not have a unique patient identifier because of conflicts between privacy advocates and those desiring an identifier that would make linkage or integration easier.

Hospital/ED discharge systems serve a variety of sponsors and customers—public health being just one of the many customers. Depending on the authority holding the administrative data

system, public health may or may not have access to data elements considered "confidential" (generally the direct identifiers of the patient if available in the system). The confidential elements are central to data linkage or integration. Without direct patient identifiers, a probabilistic methodology is required for data linkage or data integration. These more elaborate statistical methods use indirect identifiers and require additional human decision-making to assure a high quality match between data sources, and increasing costs related to the complex process in probabilistic data linkage.

States are beginning to convert to the new HIPAA standards for their data systems—primarily because of fear —fear that healthcare providers will not tolerate submission specification differences given the costs associated with their implementation. Thus, administrative data stewards are feeling the pressure to conform to the new standards, irrespective of the fact that many are not considered to be covered entities.

Other issues being faced by administrative data systems relate to the distressed financial picture in most states. While some states operate their systems based on data file fees or assessments on providers, others are wholly dependent on tax dollars. The latter systems may not receive adequate funding to allow for costs associated with data integration activities. Those relying on healthcare providers assessments may be under pressure from healthcare providers who both fund the data collection and fund the submission and correction of data. Primary attention in these systems must be on production of data files and as a result, integration may take a back seat. Those systems relying solely on fees for data may find it difficult to sell enough data to maintain the system, particularly if the HIPAA privacy regulations reduce the number of useable elements or aggregate the geography to large units that are not specific enough for market share analyses, meaning a reduction in sales of the files.

Factors Affecting Integration of Both Administrative and Clinical data

A major threat to integration is related to the quantity of information available, health care databases can be extremely large, especially in states with large populations, linking these databases can result in massive data volume. Thus, there is a compelling need to be selective about what data should be integrated. The integration should be based on rational, need-based data elements and not necessarily all the variables available for integration. There may be times when it is appropriate to include all available variables, for example, when nothing is "known" about the surveillance issue. Public health officials may need to examine correlates to better

understand new conditions or disease—this might require all the variables in the integrated files from the data sources.

Some local/state health departments report that they do not have the "horse power" in house for data integration and analysis. They have great needs for data, but do not have the platform, the expertise in data management or analysis in-house, nor do they have the capacity to handle the massive amount of data that results when clinical and administrative data systems are linked. The local public health official needs pre-aggregated data for their specific area of responsibility and they may need the expertise of other health data organizations to link this data, aggregate it, and send it to local public health entities for review.

Integration may also bring up concerns from healthcare providers, since they have submitted the information to one system and potentially were not aware of linkage plans. They may be concerned from two aspects. First, they could have concerns about the privacy of patient information, especially since the passage of HIPAA, where they are being asked to notify all patients about potential users and uses of the data. Second, healthcare providers also have concerns about being exposed, given recent reports on medical errors and other "report cards" now appearing regularly in the public domain.

Integration: Why would public health officials be interested?

Given all of the above, we must ask, "Why should public health data stewards and other administrative health data owners desire integration of their systems?" How can they be convinced of the importance of this effort given many other demands on their time?

In this section, we first discuss the reasons for integration and then give several examples of how public health surveillance systems may benefit from the addition of administrative data. Following those examples, we discuss the benefits for the data stewards of administrative data systems.

If requesters for integrated data are selective about the necessary data, data integration or substitution may improve the local/state public health entities capacity to analyze and find trends which otherwise would be hidden in volumes of data in disparate sources. Integrated administrative and clinical data can assist in case finding and monitoring of interventions. Trends that might not be found in condition-specific registries may appear in population-based data systems. If we are looking at only a registry, we might not see clusters of conditions that may be the result of environmental conditions or heredity. Missing these connections may negatively

affect our ability to prevent or intervene in the situation, to the detriment of the public. Given our lack of knowledge regarding the impact on human beings to multiple exposures to multiple chemicals, we need to continuously scan integrated data systems to locate hazards to the public.

Other reasons for integrated data relate to being able to examine the costs associated with having a specific condition, this is information that is not available in clinical databases, but is found in administrative data. We could examine costs for specific stages of disease, by having links between registry information and administrative data.

When databases are not integrated, we have less understanding of the outcomes associated with hospitalization and the various procedures that have taken place during hospital stays or in outpatient settings. Registries do not necessarily contain information on length of stay, procedures or surgical interventions and outcomes, other than death. By linking registry information to discharge data, public health officials can examine the efficacy of surgery, and other medical procedures for individuals at specific disease stages and with certain co-morbid conditions.

We provide several specific examples of the utility of integrated databases and/or substitutions for clinical databases. Again, the value of integrated data depends upon the questions that are being asked.

For example, if a public health entity has been unable to acquire a birth defects registry because of legislative opposition, it may elect to merge birth records, death records, and hospital inpatient discharge records to gain the information for monitoring trends in birth defects. A probabilistic linkage could be achieved without personal identifiers in the discharge data, by linking on such elements as: birth date, gender, date of discharge/linked to birth date, birth date to death, hospital ID, etc. However, the public health official may or may not be able to contact individuals based on this linkage, given the database design and/or policies associated with use of the administrative or vital records files. Those who can work with the limitations of administrative and vital records databases can gain information on birth defects in the population. For some questions that are population-based, this may be more valuable than looking only at registry information.

Integration of administrative data (hospital discharge) and clinical data from immunization registries can improve monitoring activities. Research has shown that immunizations can be tracked using administrative data, and that this information, is more likely to be found in the administrative data, than it is in the medical record (citation). Other questions that could be

answered relate to the number of immunization-related inpatient discharges in a specific population within a public health geographic area, and thus, public health could go beyond registration of immunizations to targeted prevention in areas with greater than expected hospitalizations. This could occur without redundant data collections, and therefore, reduce the burden on healthcare providers, while public health would still acquire the necessary information.

Emergency department data could serve as a case-finding tool for bio-terrorism activities, when public health entities and administrative stewards work together to design a real-time and population-based system. In this case, public health could collect real-time information on those "suspected cases" coming into the ED department, while the ED/hospital discharge administrative data, linked to death certificate information, could provide an opportunity for data mining to address "unsuspected" cases in the population. The real-time database could be integrated with the administrative data and this would add value not only for public health, but also for users of administrative data. Administrative data users could better understand the relationships between ED admissions and their admission complaints as well as understanding the relationship to inpatient utilization.

Another added value of data integration is the reduction of cost and burdens on the data submitters (physicians, nurses, etc.). In Wisconsin, the Bureau of Health Information has 24 data systems—a large proportion of these are data acquired directly from health care providers. BHI houses several data collections based on physician information—physicians submit data to the cancer registry, the physician office visit data collection, and vital records (birth and death). Each of these systems is idiosyncratic—the systems have varying submission due dates, data formats, electronic or paper systems, editing systems, etc. These are not the only state data systems based on physician data, other data systems in public health also collect information from physicians, these include: an immunization registry, sexually transmitted disease registry, etc… Again, these are idiosyncratic systems. Each of these data systems has specific statutory authority, each vary in terms of sponsors and constituents. What is consistent is the fact that the same physician is submitting similar data elements in different formats, via unique transmission systems, with unique rules attached to multiple sources. The authorities for these data systems must become more willing to agree to one standard for data submission, formatting, etc., before requesting additional information from healthcare providers.

In order to move beyond the idiosyncratic systems, it is clear that we must alter the conflicts between the national perspective and the states' perspective—both parties must move toward the

HIPAA standards since these are the first true standards for healthcare providers. We must convince all the various constituents of the data that while their needs may be unique, they must agree to one standard, or they will suffer more than "a flesh wound." State and Federal entities must work together, perhaps through the National Council on Legislators and the National Committee on Vital and Health Statistics, the oversight board for public health data. We must also convince the private sector administrative data collectors to adopt the same practices. Obviously, this will take considerable effort to negotiate and to change the necessary legal provisions to allow standardization and increased integration to occur.

Whether our perspective is from that of an administrative data steward or a public health authority, we must move in this direction and we must do this soon, before we are completely overwhelmed by this data collection morass. Both parties have much to lose if we don't do so. For example, if surveillance systems are rapidly deployed—policymakers, taxpayers, and health care providers may decide just to have surveillance systems—and not collect the administrative data. Alternatively, legislators could reject public health entities requests for new registries given the availability of all-payer, all patient discharge systems and the concerns about privacy of citizen information. We need both administrative and clinical data—this should not be an either/or situation.

Factors Influencing Integration of Administrative and Clinical Data for Surveillance Activities

Those interested in the integration of administrative and clinical data will face significant barriers arising from a variety of sources. Some of the barriers arise simply because administrative discharge data systems are authorized outside of state public health regulations—whether by different state statutes or by contractual relations between organizations. Other barriers will arise because of turf issues within states. These potential barriers must be understood and addressed if integration of clinical and administrative systems is desired. While the issues are complex they can be disentangled, and resolved in most cases. In Figure 1, we conceptualize the various the barriers that must be overcome to achieve integration. Each of the barriers will be discussed in the following section.

HIPAA Privacy—Effect on State Data Systems' Personally-Identifiable Elements

As mentioned earlier, some state administrative data systems reside in public health (e.g., Missouri, Minnesota, Utah, Washington, etc.), those data systems fall under the HIPAA exemption for public health data collection activities; other state administrative data systems

qualify for exemption related to questions of cost and quality of care (e.g., Wisconsin). Irrespective of acknowledged exemptions from HIPAA standards, state data systems are not immune from the impact of HIPAA privacy regulations. And, because of the variation between states' in terms of their administrative data collection practices, there is also variation in the impact of HIPAA privacy standards.

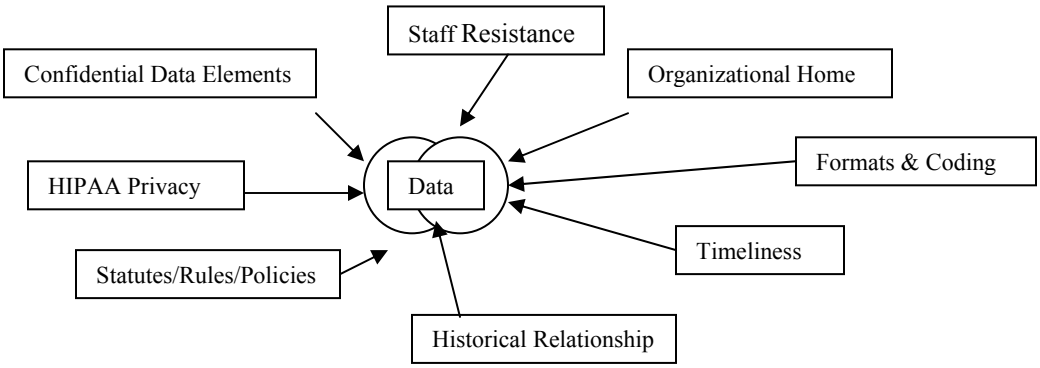

Figure 1. Barriers to Integration

The impact of HIPAA privacy may be negligible or significant, depending upon the administrative data elements collected and disseminated, the restrictions on dissemination, the ardor of providers and privacy advocates, and whether data collection is mandated, contractual or voluntary. While HIPAA does not mandate compliance for state health data organizations or public health, there may be a backlash that occurs in legislatures regarding what data elements can and cannot be collected, and which elements can be released. At least one state legislature (Wisconsin) postponed new privacy rules while they waited for the HIPAA privacy rules to become final. We may now see new state legislation that is more restrictive than HIPAA in terms of privacy of health data, which could be detrimental in terms of our ability to collect and disseminate health data. The State of Minnesota may be a harbinger of privacy activism –the Minnesota state health department has been battling with privacy advocates in and outside of government, for the last several years, as they try to seek authority to begin mandatory submission of hospital discharge data to the state. The privacy issue is "hot" even though Minnesota already has some of the most stringent privacy provisions in the nation for health research—provisions that have nearly shut down certain types of health services research in Minnesota.

In the worst case scenario, personally-identifiable elements like birth date, address, social security number, Zip code and other small geographic units, if collected, may be restricted as a result of HIPAA. In some cases, new privacy boards within the health data organizations will be

established that make determinations on "questionable" release of data elements such as: provider identifiers, race/ethnicity, Zip code. Data programs may suffer from this, given that data elements of past value (especially those for market share analysis) may not be available and, as a result, a loss of revenue will take place, putting the system at risk.

Before any data integration activities can take place, it is important to examine the impact of HIPAA on the proposed integration[3]*. It will require a thorough analysis of the necessary data elements and the potential restrictions on use whether state or federally required.*

Bureaucratic Location—A Potential Barrier to Integration

A brief description of the location of administrative databases is important for understanding the difficulties in the integration of data across administrative and clinical systems. First, identifying the *location* of the system is important in order to understand how affected the system is by the changing environment. Second, it is important to understand the differing *constituents* for administrative data associated with the location, and how constituent demands for data affect the capacity for data integration.

Administrative databases, such as hospital and emergency department discharge systems, may or may not reside within public health offices; the systems may be housed within a variety of settings, from hospital associations, quasi-public non-profits, or in state agencies other than public health. The data collection may be chartered in state statute, administrative rules, or have no state or federal authority. The steward of the system may or may not release their data to external bodies (including public health) depending upon where the authority for their existence is sited.

When the administrative data systems are co-located with the clinical data systems in a public health state agency, it is more likely that data integration will occur. However, there may still be some restrictions and barriers, given funding differences as well as statutory and administrative rules governing data use.

In planning for integration efforts, careful analysis should be undertaken to determine the impact of location on the proposed effort.

[3] The CDC has produced an excellent paper providing guidance on the HIPAA Privacy Rule and Public Health, it can be found at the following URL
http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm

Processing Timeliness and Data Editing—Potential Barriers to Linkage

Public health surveillance generally requires that submission of information from healthcare providers (and other key informants) will occur in a stream as events take place; administrative data systems generally have providers submit information in batches, monthly, quarterly or annual. Another major difference between the two types of data that affects the timeliness of the data is the data editing and cleaning process. In public health, the impact on timeliness from editing processes is minimal given the focus on surveillance. Alternatively, administrative data systems often have multiple editing and cleaning processes that take place, further delaying the release of the information. Some of the more progressive states may have administrative data available relatively soon, within a month or two after submission, other state or federal systems may take up to a year or more for release.

Linking data for integration requires selection of appropriate time periods that address the surveillance question under study. For an example, we can look at linkage time periods in the CODES project. When the selected period for the motor vehicle crash data is from January 1-December 31, 2002, the inpatient data selected for the linkage would extend 6 months beyond December 31, 2002, in order to capture charges for extended hospitalizations. Acquiring the hospital discharge data may take until December 31, 2003, or longer. While this is acceptable for examining the inpatient costs of motor vehicle accidents, this timeline may not be satisfactory for other kinds of surveillance questions related to immediate public health threats.

Early in the integration effort, a determination must be made about the appropriateness of the linkage, given issues related to timeliness of the data.

Formats, Data Definitions and Coding Systems

Unfortunately, data formats, definitions and coding systems can be a huge barrier to integration. Some databases use only three digits of the ICD-9 system while others use 5-digits in the ICD-10 coding schema. This is further complicated when 5-digit ICD-9 clinical modifications (ICD-9-CM) are used in reporting, which is common today. It should be noted that still under development is the 7-digit ICD-10 clinical modifications (ICD-10-CM) system. Some data systems have detailed code-books/users guides while others are very sketchy, making it difficult to know what you are getting. Others have re-coded the data to some idiosyncratic aggregations, making it difficult to match databases—age categories and race/ethnicity are often found in idiosyncratic categories. Some states have done extensive training with healthcare providers to assure understanding of the data definitions, yet even in those states new providers/coders take

over for employees who have left the organization and often may not understand what is meant in the data element definition. This type of error is visible only after data has been analyzed, and then only when it creates significant outliers.

Common terms may have different definitions than one might expect, terms to watch out for include: encounter, visit, anesthesia (can include both charges for hospital and anesthesiologist if employed by the hospital) emergency, and urgent care. These are often differentially defined across databases, hopefully not within. Other issues relate to the actual codes assigned to the various categories in a data element, without careful mapping of the codes, it is possible that the integrated data may contain systematic errors.

A critical examination of the details of formats, data element definitions, and the coding systems used in the databases must occur before integration. Differences should be clearly articulated to all users of the integrated database, and in any findings released from the data.

Data Collection Missions and Data Release Policies

Based on the history of the agency and its data collections, different policies regarding data release are often found in public health departments and state health data organizations. There are also different philosophies that accompany these policies, given the varying missions of the two types of organizations. Public health acquires and uses data to improve the health status of the population, and generally, does not release raw data for other purposes. Alternatively, health data organizations generally have a mandate to produce data and release it at the record level for use by others including payers, purchasers, consumers and healthcare providers.

The administrative data systems (ED and Hospital discharge) vary by state in terms of the types of data elements collected and released. Generally, state statutes or administrative rules either specifically define those elements determined to be confidential (e.g., Wisconsin) and to have restricted access, or they indicate where the authority for this decision is housed (e.g., State Division of Health). In some states the laws allow collection of individually identifiable data elements such as: name, social security number, medical record number, birth date, admission/discharge dates, Zip code, provider names, etc. In other states, many of these items are not collected and thus are not available for potential linkage or integration activities. Other states collect, but do not release health care provider identities. Some states collect and release provider names to users as long as users sign and abide with a data use agreement (Wisconsin).

One example (of many), where conflicting philosophy and policy can be problematic is when attempts are made to integrate public use cancer registry data and public use hospital discharge data. In many states, the public use hospital discharge data contains the name of the hospital as well as diagnosis codes, procedure codes and Zip codes. The cancer registry public use data does not include or allow identification of the health care provider. Even the registry data released in a de-identified and aggregated form does not allow provider identification. Yet, a data user could potentially link public use registry data with hospital discharge data and link it by county or other region, and using probabilistic matching, identify the hospitals on the integrated file given that the name of the provider is on the public use hospital discharge data. This would place the registry at risk for a violation of its own policies.

Medicaid data use agreements forbid the use of data for anything beyond what was approved in advance; this offers another set of policies that can conflict with the goal of integration. The Medicaid system has very stringent rules regarding use of Medicaid claims data; the claims data may only be used for operations of the Medicaid program, and not for other activities within a state agency—even when Medicaid and Public Health are in the same umbrella organization. Thus, if public health wanted to link Medicaid records with other administrative records it could only be done if it would be tangibly of benefit to the operations of Medicaid. For example, the Medicaid records could be used to study the impact of efforts in increasing prenatal care. By combining Medicaid utilization data and birth certificates, researchers in Wisconsin were able to study the effect of Medicaid prenatal care on the incidence of low-birth-weight infants.

We provide one last example where policy conflicts are found in how data elements are treated—that is, whether the data elements are termed confidential or sensitive. Conflicting definitions increase the complexity of data integration. In the state of Wisconsin, the Bureau of Health Information (BHI) undertook a study in year 2000, of five[4] of its main databases, to ascertain how the non-confidential, but "sensitive" data and information were defined and what restrictions were in place for these elements. The study was a more complex task than initially envisioned; it uncovered numerous conflicts between databases in terms of the elements considered to be confidential versus "sensitive" and the manner in which the non-confidential, but "sensitive" data elements were handled in terms of confidentiality and data release policies. After all database policies were reviewed, a secondary goal emerged—standardizing definitions of confidential versus "sensitive" as well as standardizing data release policies. At the conclusion of the study, even the attempt to standardize the data use agreement form across the five databases was

[4] The five data systems were moved to BHI in a merger of the Office of Health Care Information and the Center for Health Statistics (including Vital Records); each had unique histories and policies.

impossible given substantial differences in penalties for releasing information, and policies on re-release of information. The database history, funding sources, and policies for the five databases baffled the attempt to consolidate or standardize policies and forms associated with data release.

It is critical for those determining the viability of data integration to conduct a similar review of database policies to assure that output from the integration process does not result in policy violations.

Staff Resistance to Integration

Databases have stewards, who carry the responsibility for assuring the appropriate use, storage, documentation, confidentiality, and management of the data. Data stewards often develop a high degree of attachment to the database, and often resist efforts to integrate "their" data with other foreign data, to which they may or may not be given access. In essence, they feel a loss of control over the data, especially if they are locked out of the process once the data has been turned over or linked. Even the data stewards who routinely issue public use files will have concerns about re-release of the data to others following integration. To that end, some of the databases now have restrictive statutes and rules that prohibit any re-release of raw data elements even when the raw data elements have been merged with other data from additional sources.

Often resistance can be overcome by an offer to include the data steward(s) in all aspects of the project.

Another form of staff resistance is based on turf, that is, databases are part of an organizational unit, and there is a fear that integration partners may move in on their turf. This is particularly an issue when dollars become tight in an organization—staff members fear efforts to consolidate or remove the data system to the other party in the integrated data. And, often in bureaucracies, there is a tendency toward overlooking staff's database substantive knowledge, as a result moving or contracting out the data collection. The fear is based upon some history and is particularly difficult to overcome given turf issues are rarely openly discussed.

While we would like to give advice in this area, your knowledge of past practices in the organization, relationship history, and past ethical conduct in regard to "shared data" will likely be your best guide in an approach to this issue.

Other Things to Think About

Data Redundancy Issues

Though it is not cost effective to collect data more than once, it is not feasible nor desirable to completely eliminate data redundancy. We are suggesting that there is an achievable balance between appropriate redundant data collection and wasteful duplicate collection. A line in the sand, however, is that even data collected multiple times for justifiable reasons must be collected each time using the same data standards. The expense to unnecessarily translate data, as a preliminary step to facilitate needed analysis, should not be incurred. This is an unnecessary expense that drains needed funds from already very tight budgets. In better fiscal times wasteful data collection is more easily sheltered by the increased demands for data in our complex data starved world. The current fiscal realities make it imperative that we efficiently use the data that is available and carefully craft any new systems to be integrated with those existing systems when deemed appropriate.

To integrate clinical and administrative systems to better answer complex health questions data redundancy to support linkage is necessary and desirable. It is not politically or economically feasible to support a single database for answering the myriad of public health questions. The fact that there are and will continue to be multiple systems used by health officials for decision-making accentuates the need to facilitate appropriate and necessary linkage. This linkage is only possible with redundant collection of a few key variables.

The completely normalized database is pure theoretically, but often the most workable implementation alternative is to build some redundancy into the database design to improve the efficiency of the database as dictated by planned and actual use. The same is true of data collection systems used by health decision makers. The realities of when the data is needed and how is it going to be used often dictate the most practical implementation of a theoretically pure model.

Our democratic system was built on a system of series of checks and balances. It is these same principles that suggest a second important reason for some redundancy to be built into our health data system designs. With bio-terrorism threats, disease outbreaks, and community health incidents being part of our landscape, the data used to make treatment and policy decisions must be accurate. Since no single data source is 100 percent reliable, it is important that a certain

degree of redundancy be built into our system designs as a means to measure the quality of our data. The challenge for system designers is to determine where these checks and balances are necessary and where they would only add cost not value to the design. There is no single correct answer. System designers need to clearly define the critical questions and the data needed to answer those questions. It is this data that needs to be independently validated through a system of checks and balances.

Unique Personal Identifier or Data Linkage Variables?

It is possible to link files without a unique patient ID. Yet, to achieve full advantage and power from an integrated database, the data should contain a unique identifier. For example, hospital discharge data that does not contain a unique patient ID generally constrains the utilization to cross-sectional event studies. Some longitudinal outcome studies could be attempted through probabilistic linkage, but without a unique patient ID there is more uncertainty in the results of the linkage. Lack of a unique patient ID also eliminates studies of cross provider utilization by an individual. A unique patient ID should be available to assist with disease prevention, disease management, patient safety and quality of care issues. The absence of this type of linking variable also decreases the efficiency and effectiveness of linking. If we are to move to real-time surveillance activities a linking ID is critical.

We stated earlier that one justifiable case for collecting redundant data is the information necessary to provide reliable linkage between the data sets to be integrated. Typically a wide range of patient demographic information is necessary for commonly used probabilistic matching algorithms. Collection of a reliable unique personal identifier would minimize the number of data elements needed for linkage routines to function.

Because of legitimate privacy concerns, the use of any unique personal identifier should be carefully controlled and continuously monitored. Again, because of the reasonable privacy concerns surrounding the collection of a unique personal identifier, we suggest that the unique personal identifier would only be used by the data collection agencies to create other reliable linkage variables that could be released to appropriate users. The original unique patient ID should not be re-released beyond the data stewards involved in the linkage process.

The linkage variable would be a composite of enough aggregated data elements, including the unique personal identifier, to link separate data sets to create an integrated view of the data. The linkage variable that would be maintained in the file used for dissemination would be randomized and a product of de-identifying algorithms. No crosswalks between the unique personal identifier

and the linkage variable should be maintained. In this way linked records can be powerfully used in analysis without the possibility of a patient's privacy being compromised.

In summary we want to emphasize the distinction between unique personal identifiers, which should be handled with extreme care, and linkage variables that provide the power source and key to fully integrated data sets of the future.

Probabilistic Linking Variables

There are a number of states where unique patient identifiers are not available. In those states, it is critical to collect linking variables (zip code, date of birth, race/ethnicity, gender, mother's medical record number, dates, names). These linking variables should be maintained as keys for future use. Again, we must be careful to assure that the keys are carefully managed and protected from those who would use them to identify individuals and learn sensitive information. Data standards for these keys should be national standards, given the need for linking across state or other geographical and political boundaries. Standardized variables make all of our lives easier, whether data submitter, data collector, or data user. HIPAA standards provide us with a starting point for data standardization—a starting point we should embrace.

Potential Pitfalls of Integration

Impact of Integration on Sponsorship

Data systems have sponsors—those parties that pay for the collection, processing and dissemination of data and information. The sponsors may or may not want to continue their support if a new "primary" user is found. This is especially the case when the private sector is the sponsor and the state/federal government becomes the/a primary user of the data. The private sponsor may step back and expect tax dollars to cover or contribute substantially to the costs associated with collecting, processing and disseminating the information.

This potential negative impact on funding due to changing users of the data could occur for administrative data systems. The sponsor may be the provider group from whom the data is collected, and the assessments collected from the provider group may be substantial, yet this group has historically been a primary user of the data. If the primary user becomes public health, it is likely that the providers will begin to argue for a reduction in support and an increase in tax support. If not resolved, this may lead to the loss of the administrative data collection or the data

collection may be shifted to non-governmental entities governed by the providers, effectively limiting access to the data.

Responsibility for Maintenance of Effort

The integration of data will likely require substantial effort related to re-coding, cleaning, and matching the data. When only one of the parties actually use the database, the effort for the tasks or re-coding, etc., may overtax the non-user of the integrated data, and it may or may not be possible for the integration to continue. Other issues that are likely to come up include hardware and software changes in one of the systems, again issues of cost will come up at this point. If the new systems create substantial new effort and cost to maintain integration, cooperation may no longer be forthcoming.

It is important to articulate plans for maintenance of effort and shared costs for re-programming, etc., during the initial discussions. Acknowledgement of potential changes, and plans for sharing or taking on the burden will be useful down the line to prevent breakdowns.

Who is the Data Custodian?

As discussed earlier, the integrated database is made up from parts or the whole of other databases; yet, each of the databases has a custodian. Before the data is integrated, a determination of "custody" is critical, who will serve as the data custodian of the integrated database? If stewardship is split, how will day-to-day database-related activities be shared? Will the data processing be hampered or stalled by having multiple custodians? Is it possible legally to transfer custody of the data?

The "new" custodian(s) of the integrated database should be clearly articulated in the data use agreements and in statute or rule to assure that non-participating custodians are not held responsible for inappropriate use or mishandling of the integrated file.

Data integrity problems—Which custodian is responsible for data quality?

While one could argue that the responsibility for data quality falls on all data custodians, that may not be realistic if some database custodians do not have access to the final database for review and approval of the release of information from the integrated file. It is possible that during the integration process, files could be corrupted and other problems could arise in the analysis or preparation of reports. It would not be appropriate to hold the initial custodians responsible for

data quality problems that result from the integration process. However, if files have errors prior to the linkage, those errors would be the responsibility of the initial data custodian. Where this logic fails, is when the databases were designed for different purposes and, as a result, have different norms for cleaning and editing. As we've discussed, surveillance data and administrative data systems have different processes in place. Should surveillance data be more carefully edited and cleaned before merger? If so, will this slow down the process significantly, or will it be impossible due to the costs of editing clinical data, which would likely require going back to the medical record, or original paper encounter forms.

Integrated data can be "dirty" and still serve as a flag for questions—but this could become expensive and frustrating for public health field staff, when they are called upon to implement programs or deliver clinical services based on the "dirty" integrated surveillance data. Increasing editing and cleaning is expensive too. An analysis of the costs and benefits associated with editing and cleaning data and post-file analysis for errors should be undertaken, and it should be compared to the costs, financial and other, for the field staff that could result if the data has substantial errors.

The Politics of Data and Integration

Databases can be private, local, statewide, and aggregated to the federal level. At (and in) these different levels of government, there will be different political positions. If the information produced from the integrated database is not "politically correct" at all the levels involved, the findings may never surface or surface much later than is desirable due to prohibitions by the political decision-makers. When data and information are not used, then integration of the systems has been an expensive exercise, one that may jeopardize the financial support for the underlying data systems. We wanted to include two examples of situations where data became unavailable as a result of political will. When the decisions were made to pull data out of public circulation, it had a deleterious affect on important public health questions. But, we elected to avoid the "political" environment to assure that this report would be available for others to use. Your decision might need to be the same.

If report designs are approved in advance of the integration effort, there is a greater chance of avoiding situations where information is withheld, although it still does not guarantee it. Sometimes there can be agreement until the actual numbers populate the report.

Political will can and does significantly affect the amount and kind of data and information available. It should always be a consideration. Will the information from the integrated data system reveal politically sensitive issues?

Possible Systemic Outcomes of Data Integration

Motivation for the integration effort can be promoted by thinking about the bigger picture—that is, what are the gains for public health in general? And, what are the gains for the data systems if integrated?

First, if public health and administrative data stewards integrate systems, they may be able to eliminate some specialty databases or some specific data elements that are redundant, whether in the administrative data system or in public health. This will reduce the cost of collection and processing and lighten the burden for the providers who submit the data.

The integration of population based data with specialty databases is particularly useful—it allows one to look at the big picture and locate trends and determine where specific conditions appear more prevalent, and it also allows one to drill down to clinical factors in smaller areas. It is also useful to integrate data to measure program success or failure—smaller databases often cannot reflect change, or if captured small numbers may prohibit statistical analysis of the change. We need to know whether programs that are initiated make a difference in the health of the public. Measures in an integrated database could assess program effects, for example, if an education program has been in effect for management of asthma, integrated data would provide population measures of impact by assessing ER visits, inpatient stays, and mortality, as well as measures of impact on specific individuals.

Integration could also broaden the stakeholder base for public health, and for administrative data systems. This could result in support for a comprehensive national system that produces and disseminates health information for a wide variety of issues. While some are calling for a new public health information system, others have a broader vision of a new "health information system," encompassing the wider world of data users. If we do not broaden our view to the larger health information system, the same issues will re-arise around standardization and redundancy, and we will be on a collision course with other powerful interests.

Where to Start with Integration?

The principal dilemma all system designers face is how to achieve a balance between the needs for the use of the prospective data system and the capabilities of the data supplier information systems. Whatever decisions are made it is critical that data users and data suppliers maintain a trusting relationship. With this said, the starting point for integrated systems development is to listen to stakeholders describe their respective needs and capabilities. The listening discovery phase should be the starting point for all system development. It is through this listening discovery process that sustainable systems are designs are born. The purpose of all systems is to provide answers to a set of critical questions. Today's public health landscape is much more complex because of the shrinking of our world as a result of our technologies. Public health issues can no longer be viewed as a bunch of regional concerns. Outbreaks in Africa can and do effect the health of American citizens in our own communities. The answers to the critical questions will not come from a single source. No single public health information system will be capable of providing all the answers. Trying to create such mega systems will result in unsustainable solutions that ultimately would be domed to fail.

Each information system source can provide answers to only part of the public health puzzle. It is important that our system designs empower our technology to make integration possible. Framing the appropriate questions for each component of the public health infrastructure is where we should start. The greater challenge is to learn what questions need to be asked from each of the sources available to provide comprehensive answers to our most complex public health dilemmas. Involving data sources and data receivers in developing these questions is a necessary first step. Integrating these components through a shared consensus process provides us our greatest opportunities for success.

A proposed template

We have all heard the following statement: "If you have seen one Medicaid system, you have seen one Medicaid system." It is often argued that you could replace the words "Medicaid system" in the above statement with "discharge data system", "surveillance system", or "clinical laboratory system". One of the lessons learned from the HIPAA legislation is that most permutations of the statement above are not true.

The intent of the proposed template is to stimulate discussions between potential collaborating systems. By sharing basic system information we are sure there will be many opportunities to develop standard systems that can be integrated. We understand for instance that there are basic differences between our administrative and clinical surveillance systems. We are equally

convinced that there are basic similarities already shared by each of these systems. It is these similarities that make integration possible. Once systems are integrated, both would grow by the magnitude of the differences.

In the past, we believe that differences between potential partners dominated any discussion of integration. We believe the discussion should start with the similarities, while also being knowledgeable regarding differences. This template is a first cut at developing a tool to help identify these similarities and differences. The template asks some very basic questions about:

- Who pays for the data collection system
- Who uses the data
- Where is the data housed
- Under what authority is the data collected
- What is the availability of the data
- What are the key data elements
- What are the threats to the data
- What partnerships are necessary to provide the data
- What value is added by the data

The templates questions formalize the thought processes that occurred in New York State as the integrated emergency department data collection system evolved and continues to evolve.

All public health systems are faced with similar challenges, which will require dialogue and negotiation between data users, data suppliers, and others to overcome the common barriers. Creating a common tool kit is a way to nationalize the dialog that is already occurring repeatedly on a regional level.

We next provide some case examples of states at different stages in the integration of clinical and administrative data. Other states with projects similar to the two detailed below, include Maine and California.

The New York Example

A series of well-aligned stars in New York State provided an opportunity to begin development of an integrated Department of Health Emergency Department Data Collection System. On September 4, 2001, state legislation was passed mandating collection of all emergency department visits in New York State regulated hospitals. Data collection would occur through the

existing agency for the state hospital discharge system. After the events of the fall of 2001, the need for "real time" emergency department surveillance data became a high priority. The state hospital associations were very vocal that any new data collection initiatives that involved their members must be sustainable and work force neutral.

It was also equally clear to system developers that one data collection vehicle would not be sufficient to satisfy all the needs. If the only data collection vehicle satisfied the legislated mandate to collect "coded what's wrong with you and coded what's it cost (state discharge data), the data would not be timely enough to satisfy the "real time" needs of surveillance systems. If the only data collection vehicle met the need-driven "real time" surveillance systems the information available on hospital information systems within the first 24 hours of the emergency room visit would not be adequate to do disease specific research as well as comprehensive emergency room utilization analysis.

Another aspect of the "star alignment" occurred when the person given the responsibility to develop the emergency department discharge system as mandated by the legislation had an existing relationship with the person given the responsibility to develop the "real time" emergency department surveillance system. This pre-existing relationship provided the foundation for integration discussions to be part of both development initiatives. The fact that these integration discussions had occurred from the beginning of the development process for each component was a significant factor in getting industry support for each initiative.

The questions asked in the template formalize the thought processes that occurred in New York State as the integrated emergency department data collection system evolved.

Even though the legislation mandated collection of emergency department data, the state discharge system would be bounded by the capabilities of hospital information systems. That meant the data collection would need to be HIPAA compatible, which is the only system design hospital associations in New York State would support. The final design of the administrative component resulted from broad-based industry outreach initiatives. The final design strives to balance the needs of the data users with the capabilities of the data suppliers. Clearly, both stakeholders had to compromise before final agreement could be reached.

For the development of the clinical component, the limiting factor was again the capabilities of the hospital information systems. As a result of a series of separate industry outreach meetings,

it was clear the choice of available data in the first 24 hours of the visit was limited. It was also clear that to achieve industry support such a system would need to be workforce neutral to be considered sustainable by the state hospital associations. That meant the data would need to be collected electronically using business-to-business data transfer protocols.

The possibility of enriching each data system with information from the other component in the overall design elicited a great deal of excitement. From the onset both components would be developed using national standards. As stated earlier the administrative component would need to be HIPAA compatible. The clinical component was going be modeled after the Electronic Clinical Laboratory Reporting System (ECLRS), which is a NEDSS based application in New York State. Though these are two different national standards, it was easy to determine which data elements from each could facilitate linkage of the separate components into an integrated data view. The plan is that those linkage variables will be the only elements that will need to be collected in both components.

The development of each of these components of the New York State emergency department collection system is still in progress. Because HIPAA provides the base for the administrative component, the necessary hospital information system capabilities already exist across the entire state. Principally for that reason, statewide collection for the coded discharge data will begin before the fall of 2003.

The fact that electronic data available within the first 24 hours of a visit varies across the state has forced a different implementation strategy for the clinical component. Several hospitals have volunteered to participate in a pilot project. This pilot project will test the feasibility of business-to-business data transfer protocols and the usefulness of the data available in "real time" to provide the necessary surveillance alerts. Strategies for statewide data collection will be based on the results of this pilot study.

New York does not believe integration is an option any longer. It is a necessity in today's complex world. The combination of an indisputable need as a result of the events following 9/11 and present economic realities in the health care industry changes the landscape for system development. New system designs need to efficiently use available resources. All public health systems are faced with similar challenges that will require dialogue and negotiation between data users, data suppliers, and others to overcome the common barriers. Creating a common tool kit is a way to nationalize the dialog that is already occurring repeatedly on a regional level.

The Wisconsin Example

Structure and Authority for the Administrative Data Collections

In Wisconsin, the Office of Healthcare Information (OHCI) was established by the 1987 Wisconsin Act 399 (the 1998 Annual Budget Act) as a bureau level office in the Department of Health and Social Services (DHSS) and authorized to begin collecting inpatient discharge data. In addition to a governor appointed Director, the Act 399 also established the Board on Healthcare Information, a private sector policy-making board attached to DHSS. The Office of Healthcare Information was later moved to the Office of the Commissioner of Insurance in 1993, as part of a strategy for healthcare reform. It was later moved back to the Department of Health and Family Services by the 1997 Wisconsin Act 27, and housed in the Division of Public Health. Following that move, the Division of Health was split in two, to separate the Healthcare Financing activities from the Division of Health. In addition to the split, a re-organization took place that merged OHCI with the Center for Health Statistics and Vital Records, and the resulting entity was called the Bureau of Health Information (BHI). BHI was then removed from the Public Health Division and placed inside the Division of Health Care Financing.

Availability of Discharge Data to Public Health

Since 1989, public use hospital discharge databases have been available to public health, on a purchase basis. Public health has the authority (in administrative rules) to acquire the confidential data elements for their use only. No re-release of the data elements to other parties is allowed under statute. They may however, release public reports on the information, as long as no confidential elements are released.

ED Data Initiatives

An effort by a large coalition of stakeholders, to extend the discharge databases beyond inpatient and ambulatory surgery data, was successful—and the 1997 Wisconsin Act 231 allowed for the new collection of emergency department data along with physician office visit data. Historically, all databases have been funded with assessments upon the provider who supplies the data. The hospitals that were already paying for collection of inpatient data had an increase in their assessments to cover the cost of the new ED data collection. Physicians were assessed for the new office visit data collection. The latter assessment was very controversial, and as a result, a cap of $75.00 annually was the maximum the Bureau could collect from individual physicians.

The largest data user group for the inpatient data was the hospitals; the Wisconsin Hospital Association was a key supporter of the new data collections under Act 231. When all data users were surveyed by BHI regarding their interest in emergency department data, not everyone was as interested in the new data collection as were the hospitals; BHI customers expressed concerns about the potential cost of the data, and concerns about whether the data could be linked to the inpatient visits. Public health was very interested in the new emergency department data.

The emergency department administrative data will be housed in the Bureau of Health Information, given its legal mandate in Wisconsin Chapter 153, and HFS 120. It also will be under the purview of the Board on Healthcare Information. The Board has authority to determine the structure for how the data will be released, aside from specific codified limitations in statute and administrative rule.

Prior to determining which data elements would be collected, the Bureau held a technical panel on ED data, with broad representation including public health. In that meeting, healthcare providers and potential data users negotiated prospective data elements, going from the ideal to the practical, with agreement on a two-stage implementation. The first stage was based solely on data elements available on the UB-92; the second stage of data elements included clinical data elements.

While the collected data elements will be based on the UB-92, they are subject to abstraction rather than a copy of the UB-92. This was the preference of hospitals, given their desire to maintain comparability with the system that is in place for the discharge data collection. Hospitals are required to recode information related to payers, to assure that individuals cannot be identified by their plan. Patient name, address and other direct identifiers are not submitted to BHI, as a result, there is no unique patient identifier that allows tracking across institutions in the database, however, it is possible to link ED and inpatient care with a facility via an encrypted case ID assigned by the facility. The key ED data elements include:

- Facility ID
- Patient Control Number
- Patient Medical Record Number
- ED Discharge Date
- Patient Home Zip code
- Patient Date of Birth
- Patient Gender

- ED Admission Date
- ED Admission Source
- ED Discharge Status
- Adjusted Total Charges
- Primary and Secondary Payer Type
- Diagnosis Codes
- E-Codes
- Procedure Codes
- Attending Physician ID
- Other physician ID
- Type of Bill
- Encrypted Case ID

Clinical elements that are in phase two will require new administrative rules. Wisconsin Ch. 53, Stats., requires specification of the individual data elements in the administrative rules. Thus, the change to Phase two data elements will require new rules; administrative rules on average, take approximately one year, from start to finish to complete.

The first phase of the data collection began first quarter of 2003. Given the experience of the hospitals in submitting discharge data, the data will be available for release as soon as it is processed. A decision has been made to provide the first quarter of data at no charge to the purchasers of the inpatient data, to give them a chance to use the data and determine its value. While public health anticipates the new administrative data, it is clear that additional data will be needed for surveillance activities related to Bio-Terrorism. Those data elements will fall under the authority of public health. Some preliminary discussions have taken place regarding the integration of the Public Health clinical data elements and the BHI administrative data elements.

The ED data will be processed similarly to the inpatient discharge data. Data is submitted on a quarterly basis, from the hospital to the Bureau via an electronic system of submission. The system has automated data edits and profiles that are returned to the provider on a private bulletin board for correction or verification. There are also a variety of quality assurance activities that take place prior to release of the data. Processing time is approximately 90 -120 days.

A Lucky Break for Integration

Fortuitously, the Division of Public Health Administrator for eight years has now been named the Director of BHI. This should facilitate greater dialogue between the Division of Public Health

and BHI. In addition, with Bio-Terrorism and HANS financial support, Public Health has created a new web-based system that has the potential to serve as a single portal for all of Public Health and BHI data. Some of the burden that exists for healthcare providers submitting nearly the same information to multiple sources would be reduced, by having a single portal for both public health data and administrative data. This would be advantageous to all parties, as cooperation would be enhanced, and data collection costs reduced. Health care providers who are charged assessments to maintain administrative systems would be spared the additional cost of programming for two separate idiosyncratic submission systems. State and Federal tax dollars could also be saved, by using a single portal for submission, otherwise dual systems would need to be in place.

Threats to Integration

There are a number of threats to data integration, including those related to the financing mechanisms of the systems. While Public Health has primarily Federal and State tax support, BHI receives assessment dollars and program revenue for the administrative data collections. The Board on Healthcare Information has representatives of the providers as members; if the Board believes that Public Health will be the primary user of the ED data, they will push towards a reduction in their support of the data collection. In the past, they have publicly discussed this, and it is likely they would go forward to the legislature, with demands for relief from the assessment.

In the details of this ED data collection, exists the potential for data quality problems. Caution should be used in regard to co-morbidities, complications, and diagnostic coding. Integration with Public Health clinical elements that are captured by nursing or medical staff may differ from the information that is coded by hospital coders for the UB-92 from the medical record. Quality assurance activities will need to attend to potential incongruence between the clinical and the billing information on the file.

Another potential problem is identifying the inpatient cases that started in the ED, because the source of admission may not always be accurate in the inpatient data, making it difficult to assess whether the patient was transferred in from the ED. For example, if the question to be asked of the data relates to the process of care for CHF, it will be important to know how many cases of CHF came through to inpatient from the ED, and how many CHF cases used only ED services, and how many were referred directly to inpatient from their physician.

Another potential threat to the data integration may come from privacy advocates. There are individuals and groups that could pursue changes to the statutory language that would prohibit data linkage and thus, data integration, based on fear an individual's privacy may be threatened

by integrated data systems. An equally real threat is that data is not used to its full potential to avoid potential confrontation with privacy advocates.

It is also possible that staff on both sides may not cooperate as fully as necessary given that both parties will be collecting emergency department data, and this portends to some turf protectionism. Given the significant effort by BHI staff to assist in the passing of legislation and administrative rules, it may be difficult to convince them of either the single portal plan or the integration of the data. Bringing forward statutory or administrative rule changes into the legislature essentially re-opens the discussion on the entire set of data collections, a potentially hazardous action, given there are still legislators interested in ending, or further restricting the state health data collections.

Value of an Integrated Administrative and Clinical ED File for Public Health in Wisconsin

Integration of clinical and administrative ED data allows one to move from disease or event specific data to population wide data—extending the potential range of analysis. For example, without integrated clinical and administrative ED data, it may be difficult to make accurate estimates of the number and costs of injuries occurring to children or adults (up to age 65). Value is added to MA recipient data (using the ED) by adding the other payer information; the integrated ED data can be used for a variety of programmatic needs, such as program evaluation of prevention efforts, or for comparing utilization and access to ED care across payer types, etc.

In terms of Bio-Terrorism, having an integrated ED data system means that rapid case finding can occur through either real-time clinical symptom access or via data-mining to detect clusters of symptoms leading to earlier identification. Clusters of cases that cross over registry boundaries or fall outside the registry area can be located and interventions can be planned for that area.

Clearly, there will be additional benefits of integration—benefits that cannot be envisioned and articulated in advance.

Attributes of Successful Integration Efforts

Our vision of a successful integration effort includes the following attributes:

- All parties to the data integration process are knowledgeable about the history and culture of the databases to be integrated, the constituents for each database, the restrictions on use, the financing mechanisms, the statutory and administrative rules, and the release policies

- There are clearly stated questions to be answered through the integration of data
- There has been an agreement to the custodial relationships and to the process of approved releases
- Maintenance of the resource has been planned (if appropriate)
- Necessary statutory and administrative rule changes have been completed
- An analysis has been completed in relation to issues of personal identification in the integrated database
- A data-use form has been agreed upon
- Detailed storage and security plans have been approved by all parties

Conclusions

In order to advance the capacity of our current data systems to answer current and future questions, we must continually assess how and when the public health clinical systems and the administrative data systems can be integrated. This paper has provided an introductory discussion of the value of linkage and recommendations for a process to address known barriers to linkage. While we believe these strategies will assist the data custodians in the process of linkage, custodians must first acknowledge the myriad forces at work, and second, be willing to stick with the process to its conclusion.

We also recommended integration efforts should be based on strategic goals—to meet specific needs—not the establishment of large data warehouses for some unknown need. Instead of building large warehouses, we can address unknown needs by standardizing data elements, thus, readying the data for strategic linkage. Strategic linkage is part of system design from the onset, and is not just a good idea implemented retrospectively.

While integration efforts can take place without a unique patient ID, we strongly suggest movement to a unique patient ID; as a society we must overcome our fears by establishing mechanisms that safeguard our privacy while improving our capacity to understand disease, improve treatment outcomes, and protect us from terrorist, biologic, and environmental threats.

We also recommend that efforts should be expended to assist data custodians with de-identification processes. If we do not do this, valuable data will never be released for use.

And we encourage efforts to speed up the data collection, cleaning, linkage, and dissemination processes. To answer our questions today, we need data to be available in a more timely fashion. Data has a short shelf life.

In conclusion—we know that we don't have all the answers—in some cases we haven't formed the right questions. We ask that readers of this document share their thoughts and experiences with the National Association of Health Data Organizations, which in turn can disseminate new ideas via a listserv, conferences, and pilot projects.