Public Health Data Dissemination Guidelines: *NAHDO Working Technical Paper Series*

Funded by the Centers for Disease Control
Assessment Initiative

July 2005

Edited By: Gulzar H. Shah, MStat, MSS, Ph.D.

The National Association of
Health Data Organizations

ACKNOWLEDGEMENTS

The National Association of Health Data Organizations (NAHDO) developed this document with funding and support from the Centers for Disease Control and Prevention (CDC) Assessment Initiative. We are grateful for the leadership and guidance provided by the CDC Project Team, Patricia Schumacher and Alex Charleston.

The lead author of this document is Barbara Rudolph, Ph.D., NAHDO's Data and Research Senior Scientist Consultant. Dr. Rudolph integrated documents and research conducted by Michael Stoto, Ph.D., Harvard University; and Luis Paita, Ph.D., NAHDO's former Deputy Director. Guidance for this paper was provided by the NAHDO-NAPHSIS Expert Panel consisting of Garland Land, Missouri Department of Health and Human Services; Lois Haggard, Ph.D., Utah Department of Health; David Solet, Ph.D., King County Health Department; Gregory Crawford, Kansas Department of Health and Environment; and Daniel Goldman, M.D., M.P.H., Vital Net. We thank the Assessment Initiative states and the private and public state health data organization staff who selflessly provided input and shared their data dissemination practices with us.

Gulzar Shah, Ph.D, Director of Research at NAHDO authored the Small Areas section of this paper.

The NAHDO-CDC Cooperative Agreement is a joint project promoting collaboration and partnership with the National Association of Public Health Statistics and Information Systems (NAPHSIS). As national associations representing statewide population-based health data bases and the agencies that maintain them, NAHDO and NAPHSIS, with support from CDC, are working to build data and technical capacity at the state and local levels.

Guideline Use

We hope users of this document will notify the National Association of Health Data Organizations with additions or corrections. Please send an email to: Gulzar Shah, Ph.D. gshah@nahdo.org.

Table of Contents

Statistical Approaches for Small Numbers: Addressing Reliability and Disclosure Risk

NAHDO-CDC Cooperative Agreement Project
CDC Assessment Initiative

By: Luis Paita, PhD.
Barbara Rudolph, PhD
Gulzar H. Shah, MStat, MSS, Ph.D.

The National Association of
Health Data Organizations

Statistical Approaches for Small Numbers: Addressing Reliability and Disclosure Risk

A. Introduction

Public health data when queried or displayed in web-based data tables can often have cells with a small number of individuals or events especially when the query is focused on small geographic areas (Zip codes), rare events, population subgroups, provider groups, payers, or other small samples. The primary statistical concern is reliability of results from queries in which the results contain small cell sizes or a small underlying population. Without some intervention to increase cell size or population, there may be misinterpretation by the user. It should be noted however, that the definition of "small" varies across political boundaries, the databases and states. The application often influence how the term "small" is defined.

Most public health professionals are aware of reliability problems resulting from small numerators; fewer consider a small denominator as a contributor to poor reliability of the data. Web-based data systems' developers should be aware that the reliability of rates based on case reports where the denominator is from a smaller population will be affected negatively. "For example, if x forms the numerator of a rate p, population = n, when p is small Var (p) = Var (x/n) = p/n, the resulting standard deviation for the rate is significantly larger in smaller populations. In the table below we see that a denominator with 100 cases results in a less reliable rate than one with 10,000 cases where both have the same case numerator." (Stoto, 2002:18)

x = 4,	**n = 100 or 10,000**	
x = 4	n = 100	SD(p) = √0.04/100 = 0.02
x = 4	n = 10,000	SD(p) = √0.0004/10,000 = 0.0002

p= 0.04	**n = 100 or 10,000**	
p = 0.04,	n = 100 or 0,000	SD(p) = √0.04/100 = 0.02
p = 0.04	n = 10,000	SD(p) = √0.04/10,000 = 0.002A

Adopted from Michael Stoto, 2002[1]

While small cell size is a concern for most public health statistical publications, it is more acutely so in web-based data dissemination systems for several reasons. First, because web-based data dissemination systems are particularly desirable for immediate answers to questions about the public's health, and generally, the users of the systems are interested in data for small geographical areas and other small groups of individuals. Second, the information reaches a much broader audience than a paper publication, and often this includes individuals without statistical or epidemiologic training. Third, web-based systems generally provide less documentation on how to interpret the results than do paper publications which usually provide extensive bibliographies, appendices, footnotes, caveats, etc. The web-based data dissemination systems (WDDS) that attempt to provide documentation still have the issue of varying types of queries—it may be difficult to direct the user to the appropriate documentation.

Small numbers are also of great concern when reporting sensitive information that might lead to violation of individuals' right to anonymity and privacy with respect to attributes that are typically stigmatized. While this guideline is primarily focused on data reliability, it also provides statistical approaches that can better assure anonymity and privacy.

This document is part of a set of guidelines supported by the CDC Assessment Initiative and designed to assist data managers, epidemiologists, and analysts in public health when releasing public health data on the web. These guideline sets include: Statistical Approaches for Small Numbers**:** Addressing Reliability and Disclosure Risk, Security of Data for Web-based Data Dissemination Tools, and Management and Institutional Controls for Reducing Disclosure Risk in Web-based Data Dissemination of Public Health Data; as a package, the guideline set will assist in assuring reduction of risk of inappropriate disclosure of sensitive information, meaningful statistics, and security of data. We have attempted in the guideline set to artificially isolate methods to reduce redundancy across the guideline set, however, in practice one would use methods from each of the guidelines.

This specific guideline will address statistical approaches for releasing public health data on web-based dissemination systems; approaches that impact on the reliability of statistical tests and/or, at the same time protect individuals from disclosure of sensitive information.[2] The first part of the document covers methods operating as modifiers of the data in the underlying database—the section is titled "Summary of Data Modification Statistical Approaches for Addressing Small Cell Size. First, there is a summary of the approaches, followed by individual sections discussing in greater detail each approach. The second major section is titled "Summary of Formal Statistical Approaches to Improve Interpretation of Results from Small Cells." This section includes statistics for modifying the interpretation of the results of statistical tests. It too, has a brief summary, following by descriptions of the individual approaches. Within each description we offer references, web-sites where the approach has been used and contact information.

[2] Non-statistical approaches for reducing disclosure risk are found in the "Management and Institutional Controls for Reducing Disclosure Risk in Web-based Data Dissemination of Public Health Data" guideline.

B. Summary of Data Modification Statistical Approaches for Addressing Small Cell Size

There are a variety of approaches for increasing the reliability of statistical tests in situations where cell sizes are small; these statistical techniques address the issue with both modifications to existing data and the use of synthetic information to achieve larger cell or population sizes.

The "data modification" statistical approaches include the following:

Aggregation or combining results over geographic areas, or multiple years, or subgroups (e.g., age groups) is done in order to achieve a larger denominator that will produce a larger cell size in the table.

Statistical Noise/Data Perturbation. Introduction of uncertainty to all cell values in a table less than a pre-determined threshold (e.g., < 10 observations). To implement one can add or multiply the values of a continuous data element by a randomly-determined factor.

Smoothing Techniques--including maximum likelihood, simple weighted averages, and the moments methods (multivariate signal extraction) are all classified as approaches for smoothing or signal extraction. These are designed to improve the reliability of the estimates by removing noise from the data.

Other Bayesian Methods--small area model-based estimation and bootstrapping. Essentially these techniques impute information from either direct or indirect sources to create a new estimate for the geographic area, or for a specific demographic characteristic. Bayesian methods are primarily used for improving data reliability, but also serve to reduce disclosure risk.

Aggregation of results is the most commonly used statistical approaches to address small cell sizes; the other approaches are more complex and require more statistical sophistication to implement and potentially more time to compute. This latter point will impact a dynamic web-based query system—users will wait only seconds, rather than minutes or hours for computations to occur.

We will describe in greater detail each of the data modification and interpretation approaches—indicating their strengths and weaknesses. Also provided are public health agencies/systems that have utilized the method described. In addition, we provide the user with useful references for further investigation.

C. Data Modification Approaches

Aggregation

This approach combines results over geographic areas, or multiple years, or subgroups (e.g., age groups) in order to achieve a minimum number in the combined cell. For example, if the results for a 5 year age group (ages 1-5 years of age) do not yield an adequate number of cases for statistical testing, then the age group is extended to cover more ages (1-10 years of age). This however, results in loss of information and precludes an analyst from drawing any conclusions about the children age 5 and under. In rural areas, it is very difficult to construct grants or planning documents for specific health care conditions, due to small numbers in cells. One cannot make reliable statements for 1 or 2 cases of cancer, or conditions potentially resulting from environmental factors. So, to achieve that reliability we aggregate, but there is a loss of detail. If aggregation isn't used, then results are often blocked from display via cell suppression techniques and the utility of the data is reduced significantly. These cell suppression methods are described in greater detail in the guideline "Management and Institutional Controls for Reducing Disclosure Risk in Web-based Data Dissemination of Public Health Data".

The actual number of cases required for statistical reliability may vary by type of statistical test used, by the level of significance (alpha level), and whether confidence intervals or coefficients of variation are produced. The other issue is how the results are going to be used, that is, if the results are for exploratory investigation versus determining how millions of community dollars will be spent. Clearly, the more important the results are—the greater the reliability that is required.

Agencies Using This Approach

It is likely that all web-based systems have been designed to include some aggregation of results.

Strengths and Limitations of Approach

This approach results in loss of information, e.g., aggregating data for two or more entities may hide differences between or among those entities. Yet, aggregating data for multiple years provides more stable estimates—estimates that are less likely to be influenced by random variation. But, the data are consequently older when aggregating across years.

While aggregation is relatively easy to do, often in dynamic web-based query systems it still may take the user a number of tries to select age groups or geographic units to achieve the cell size that is necessary. The return of information is generally limited by suppression algorithms to avoid any disclosure.

Some systems are designed to avoid this by pre-aggregations within the database. This may at times unnecessarily reduce information for more common events or conditions. It does, however, assure that you will actually receive information in cells for statistical testing. For example, a web-based data dissemination system could be designed to pre-aggregate age groups to reach a minimum of 30 individuals in a cell. This would require advance programming; and perhaps, loss of information for certain types of questions in the upper and lower ends of age categories.

Standard Data Perturbation Methods – Statistical Noise, Data Swapping and Controlled Rounding to Reduce Disclosure Risk

These Bayesian methods include the addition of statistical noise, data swapping and controlled rounding. The addition of statistical noise to a data file, data swapping or controlled rounding results in "pseudo" information that reduces disclosure risk. Other Bayesian methods described in the next section use synthetic data to increase the size of small cells, thereby increasing data reliability.

There are two types of “noise” that can be introduced into data, natural noise and statistical noise (Zarate, 1999). Natural noise consists of errors in the data, such as: coding errors, keying errors, or missing data. Statistical noise is introduced into the data to add uncertainty to all cell values in a table that are less than a prescribed threshold, such as < 10. For instance, when the results return 4 cases and the threshold is 10, an addition is made to some case values in the table. This results in blurring of the data, assuring that identifying an individual within the data is protected to a degree of certainty. While the blurring does change the data, the simple statistics and distributions remain the same. The add-on factor is not made available to data users.

“Controlled tabular adjustment” (CTA) is another form of statistical noise—which can be used in two-dimensional tables to reduce disclosure risk, as an alternative to cell suppression. Unfortunately, it cannot be extended to tables with three or more dimensions (Ernst 1989). It relies on a probability measure for rounding “down” or “up” for each of the table cells and uses a mathematical programming approach called a *stepping stones algorithm*. Using an “unbiased” controlled rounding approach can preserve original values with respect to the statistical criterion expectation. The results deliver the same statistical distribution, assuring reduction of disclosure risk (Cox, 1987). The National Center for Health Statistics has funded the development of software for tabular data protection using controlled rounding and a method to preserve additivity of the sub-totals along one of the dimensions (rows or columns.) The software uses a synthetic substitution for replacing a cell value; it substitutes the current value of the cell with its “closest safe value” and uses linear programming to adjust other cells to preserve additivity (Gonzalez and Cox 2004).

“Data swapping” is a method that swaps information from one individual within the same sample to another individual with similar characteristics in the sample. This results in “pseudo-cases.” The individual records (after swapping) do not represent any one individual, yet these pseudo cases still produce the same simple statistics and distributions as those produced by the original data. This allows for display of small cell sizes without risk of individual identification.

An example of data swapping can be found in the Census Bureau’s “confidentiality edit.” The Census Bureau developed the “confidentiality edit” (CE) to prevent the disclosure of personal data in tabular presentations. The CE selects a small sample of cases and interchanges their data with other cases which have same characteristics on a pre-selected set of variables but who live in different geographic locations (Jabine, 1993).

One survey currently uses multiple imputation methods; the Survey of Consumer Finances, which is conducted by the Federal Reserve Board and holds sensitive financial information from a high-wealth population, has been a test bed for this type of method. This method is computationally intense, and is not likely to be applied in any web-based data dissemination system that allows for dynamic queries.

Any of these standard data perturbation methods add expense to the preparation of a file for web-based data dissemination, but for web-based micro-data files, this method can be useful in preventing users from matching the database with other databases for explicitly identifying the individuals in the second database. Thus, the perturbation methods are useful for protecting confidentiality of the data, as well as increasing data reliability.

Agencies/Systems Using This Approach

U.S. Census Bureau

National Center for Health Statistics

Strengths and Weaknesses of the Approach

Perturbation methods such as the addition of statistical noise, data swapping, and controlled rounding limit disclosure risk while maximizing information available to the user. Although it may distort the actual information it maintains the statistical distribution. The distortion however, may result in misinterpretation by users and produce unnecessary concern about specific health conditions or environmental risks—when examining cell sizes that are "rounded up."

Should perturbation methods (if less computationally intense) be applied to data in a web-based data dissemination system? It appears that they may work for some simple statistics, but could be problematic depending on the use of the results. Given that many social scientists are skeptical of analyses that are not based on the original data, the perturbation methods should be applied last, when other approaches cannot prevent disclosure (National Research Council, 2000). And if such methods are used, there must be greater effort to education and convince the data users that the key properties of the data are preserved, even with the addition of statistical noise (imputation).

Data Smoothing for Improving Reliability

Data smoothing is a technique that adjusts for differences in the reliability of data resulting from small cell sizes. The inferences bases on tabular cells that hold small numbers are less reliable than those based on relatively larger numbers.. There are a variety of approaches to producing smoother data, including maximum likelihood, simple weighted averages, and the moments methods (multivariate signal extraction). They can all be classified as approaches for "smoothing" or signal extraction.

Geographic Smoothing Methods

Smoothing is a widely used technique to adjust for differences in reliability of data associated with small numbers. Smoothing refers to removing the smaller random fluctuations resulting from random errors, by averaging data in space or time in order to see the trends (Simonoff,1996). Geographic smoothing techniques have often been used in conjunction with the creation of disease incidence rate maps, where the raw rates for rare events (such as cancers) are unstable for regions with small populations at risk. Rather than report small numbers for a specific geographic region (Zip code or census block), typically disease incidence has been reported only as a summary count or rates for a larger defined region, such as county. Geographic smoothing techniques can be used to produce counts for smaller geographic regions with small numbers at risk. The statistical models draw upon the "strength offered by adjacent geographic areas" to create more stable estimates for the small area. The smoothing methods may also rely upon Bayesian or empirical Bayesian modeling approaches described below.

While it has not yet become normative for public health systems to utilize smoothing techniques in their web-based data dissemination systems, there are several systems that do currently rely on geographic smoothing methods.

Recently, HCUPnet has utilized geographic smoothing in their risk-adjustment methodology for reporting hospital indicators. While not the same usage, this methodology was again used with the idea that small cell sizes could be made more reliable during the development of the risk-adjustment methodology. There is a report available on this technique on the AHRQ website (www.ahrq.gov) under HCUP.

The State of Washington's EPI QMS system uses both smoothing and Bayesian methods for addressing small cell sizes. These included adding additional data from other years, or drawing on data from surrounding "neighborhoods." This system is designed primarily for epidemiologists and not the lay public, although some data are available to the public.

Agencies/Systems Using Geographic Smoothing Approach

Utah Department of Health, Indicator Based Information System (IBIS) at
http://ibis/health/utah.gov
Contact: Lois Haggard, Ph.D., Utah Department of Health
loishaggard@utah.gov

State of Washington--Epidemiologic Query and Mapping System (EPI QMS)
https://fortress.wa.gov/doh/epiqms

State of Washington—VISTA system
GIS and Spatial Epidemiology

www.doh.wa.gov/OS/Vista/HOMEPAGE.HTM
Contacts: Dick Hoskins
reh0303@hub.doh.wa.gov
360-236-4270

AHRQ, HCUPnet http://www.ahrq.gov/data/hcup/
Uses smoothing techniques in the QIs.

Bayesian Methods[3] for Improving Reliability

Statistical procedures have been developed to address small numbers in sample data. These procedures draw upon Bayesian methods and include small area estimation (see for example, Shen and Louis, 1999 & 2000). Essentially, these techniques impute information from either direct or indirect sources to create a new "estimate" for the geographic area, or for a specific demographic characteristic. The techniques use information from other sources or from population and cell averages, replacing the sample data by using only the new information, or by averaging the new information with the sample information. The latter, is called a composite estimate (Ghosh and Rao, 1994). It creates new estimates that are based on the mean of the sample data and the other external source's mean. Sometimes the estimates are also weighted. Using these methods assumes that the other source of information is an equal or better representation

[3] Bayesian statistics rely heavily on the formulation that "posterior is proportional to prior times likelihood." This translates to—the basis of various alternative hypotheses is knowledge at a particular point in time, modifying those hypotheses is based on collecting new information from relevant data to arrive at "posterior probabilities," essentially being able to predict both sensitivity and specificity of the estimates.

of the population than the sample. While sampling statisticians have used these techniques for some time, they have not been widely used in public health web-based data dissemination systems, because calculating the variance for these techniques is quite complex. For model specifications, see the Ghosh and Rao, 1994.

Bayesian modeling approaches are used to address the reliability of data (given small cell sizes) and to predict a better estimate using information from prior quarters or years of data or other sources, thereby reducing the variability or error from estimations of the small cell value based on using only the population mean. Census Bureau researchers (Fay and Herriot, 1979) used Bayesian methods with census data; they proposed that an estimate of per capita income (PCI) for a small place in the census could be a weighted average of the census sample estimate and a "synthetic" estimate obtained by fitting a linear regression equation to the sample estimates of PCI, using other data sources for the independent variables, such as county averages and tax-return data. The Census Bureau adopted this approach for estimates of PCI in small areas in 1974. The National Center for Health Statistics also adopted this synthetic approach for creating state estimates of disability for the National Health Interview Survey data.

A Bayesian modeling approach could be used for small cell size estimates in hospital discharge data reporting by taking information from previous years for the variable of interest, and using this synthetic estimate along with a weighted average of the current year. Estimation can occur using a fairly general Bayesian regression model. However, the Bayesian methods may pose a challenge for dynamic web-based dissemination systems, given the computational time for these estimations. And, in some cases where normality is violated, the models may not assign the appropriate weights. In order to assess whether the Bayesian method used is appropriate, various regression diagnostics may be necessary. Software, such as: LISREL 8.0[4] allows an assessment of the reliability of the prediction and thus could provide a test of whether reliability was increased using Bayesian methods.[5] But, integrating software like this would be difficult in a dynamic web-based system.

Bootstrapping Approaches

Bootstrapping approaches for estimating various parameters from the sample for the purpose of studying the mean and variance of these parameter estimates can be used for ascertaining a reliable estimate of a small cell in a table. Monte Carlo techniques (a form of bootstrapping) are essentially computer-generated data based upon the available sample. There are a variety of software packages that provide for this approach. Bootstrapping allows you to produce estimates of standard errors by repeated random sampling (with replacement) from the available sample.(Vgot, 1993). Generally, users of this technique will draw anywhere from 100 to 1000 sub-samples from the existing data to generate the estimates. While the procedure itself is not as complicated as some of the

[4] K. Joreskog, D. Sorbom, S. duToit, and M. duToit. (2000) LISREL 8: New Statistical Features, Scientific Software International, Lincolnwood, Illinois.

[5] If you want to predict the contents of a small cell, predictions can be estimated using general Bayesian regression models. Alternative Bayesian approaches use covariance matrices and the likelihood function in multilevel models, where the actual value is the fixed part of the model, the random component is the estimated population parameter, or predicted cell count. With this approach variances from the estimated or predicted cell count can be estimated using software such as: LISREL 8.0, this software allows an assessment of the reliability of the estimated prediction.

other Bayesian methods, it does require substantial computing resources. For example, Waller et al., (1997) stated that a model using bootstrapping (with 500 iterations) to estimate disease incidence in a geographic area took 20 minutes on a Sparc10 workstation, thus this approach may be untenable for dynamic web-based data dissemination systems, where responses to queries should not take longer than 1-2 minutes. This could be resolved, however, with faster computers, and a static response (to the dynamic query) based on pre-aggregated tables.

Strengths and Weaknesses of Bayesian Approaches

Critics of these methods suggest that the end users may assume that data created from iterative statistical sampling processes is the true data, which may result in misinterpretations. Critics are also concerned that the new estimate may distort the true relationships in more complex modeling efforts.

In addition to criticism pertaining to data integrity, there is also an implementation challenge in terms of web-based systems. The computer processing time is significant when creating the new cell estimates. It is unlikely that a dynamic or "query-on-the-fly" system could utilize this method, given the processing time demands.

These more advanced methods require highly trained public health data managers, IT staff, and analysts. Alternatively, financing is needed for contracts to build the systems that incorporate these more complex methods. In addition, data users will need training in interpretation of the query results. Sophisticated, but easy to understand documentation, will be required for users.

D. Summary of Formal Statistical Approaches to Improve Interpretation of Results from Small Cells

Formal statistical approaches (confidence intervals, hypothesis tests χ^2, and coefficients of variation) allow maximum information to be disseminated while honestly communicating statistical reliability to the user. The evaluation of reliability of any measurement procedure consists in determining how much of the variation in scores among individuals is due to inconsistencies in measurement.(Seltiz et.al.1967). If measurement is free from random errors, it is considered reliable. Some measures of reliability focus on measuring different sources of variation. We describe three formal statistical approaches to improving the quality of interpretation of web-based query system results.

Calculation of confidence intervals is a strategy to provide the end user with a more accurate interpretation of the results. The width of the confidence interval provides a good picture of the potential variability in the results that is attributable to random error. The smaller the number of cases in the numerator and denominator the greater is the width of the confidence interval. When comparing an estimate E_a (rate, ratio, mean etc.) with another more stable estimate E_b (e.g., state or national rates), if the estimate being compared (E_b) falls outside of the confidence interval of the $E_{a,}$ one can be confident that it is not the result of randomness, and that they are statistically significantly different.

χ^2 tests are used when the distribution of the data is not normal, and the data are not at least at a level of interval scaling. These are considered non-parametric tests because they fail to meet the basic parameters as described necessary for parametric tests of normal distributions.   Certain types of chi square tests are designed specifically to adjust for small cell size.

Coefficient of variation (C.V.) is a measure of the variation around the estimate and thus the stability of the estimate, compared with the magnitude of the estimate.

In remainder of the chapter, we further describe each of the above three statistical approaches that can be used with small cell sizes for increasing the understanding of the results. These methods are available to designers of web-based data dissemination tools. It should be remembered however, that many individuals will not understand how they work or what they mean.

Confidence Intervals

When dealing with health data containing small cell sized in tabular data, wide fluctuations in the data can occur from quarter-to-quarter, or year-to-year, when the cell size is small. For example, if examining hospital discharge data, it is likely that some hospitals will experience very small numbers of deaths for certain types of diagnoses or procedures, and the number of procedures may also vary across institutions. Analysis of changes from one year to the next could result in significant findings, yet these findings could be due solely to chance. Using a confidence interval (C.I.) assists the data user in determining the reliability of the information being provided. The wider the confidence interval the less precision there is in the estimate. Narrow confidence intervals suggest that the estimate is nearly precise, and that chance plays a smaller role in the outcome of interest.

Confidence intervals avoid many of the problems inherent in simply reporting the statistical significance of a test statistic, and provide considerably more information. A significant statistic only gives us the information that the sample statistic is different (or not different) from the true population parameter; or two sample statistics are different (or not different) from each other. Confidence intervals, focus on the magnitude of the difference, and provide an estimate on the precision with which the population parameter is estimated.

When to use confidence intervals? Oakes (1986) suggests that confidence intervals be used whenever there is a need to understand the uncertainty in a point estimate. That uncertainty often arises due to small cell sizes. In hospital reporting, if there is significant variation across time in the hospital’s mortality or length of stay, or average charges—a confidence interval will help the user to understand the contribution of chance variation to that fluctuation.   The analysis of fluctuations across time (for institutions whose population is smaller) using confidence intervals will allow a user to differentiate actual change over time from change caused by the instability of rates due to small numbers. Accordingly, it can also help to reduce the misinterpretation of random variation when cells are small. Institutions that have large numbers of cases for cells will have narrow confidence intervals, suggesting that, there is greater precision and smaller margin for error in interpreting the outcomes.

In addition, confidence intervals can be constructed around an odds ratio for different levels of error (alpha), an important contribution to epidemiological research. For example, reporting on the odds of death in a particular hospital for a particular diagnosis or procedure is of interest to consumers. It is important that users not compare mortality rates across institutions by examining overlapping confidence intervals.

To create a confidence interval requires only an approximate normal distribution and knowledge of the standard error of the sample.

The State of Washington Department of Health has provided excellent documentation of the methods by which they produce confidence intervals for their web-based data systems. They provide the methods used for producing confidence intervals for the following: age-adjusted rates, crude and age specific rates, standardized mortality rates (for cells with <100 and >100 cases), and for non-independence of events (such as those including multiple re-admissions), binomial proportions and for complex survey design. See the web address below to download a PDF version of their guidelines.

Agencies/Systems Using Approach

State of Washington, Department of Health
Documentation available on website (generally designed for public health professional use) www.doh.wa.gov/data/guidelines/
Vista/PHw - Washington State Center for Health Statistics
Contact: David Solet, Washington Center for Health Statistics.
(David.Solet@doh.wa.gov)

Utah Department of Health, Indicator Based Information System (IBIS)
Contact: Lois Haggard, Ph.D., Utah Department of Health
loishaggard@utah.gov

State of Wisconsin
Bureau of Health Information and Policy
Wisconsin Interactive Statistics on Health (WISH)
www.dhfs.state.wi.us/wish
Contact: Richard Miller (millere1@dhfs.state.wi.us)

Missouri Department of Health and Senior Services
Missouri Information for Community Assessment (MICA), (C.I. for Rates)
Contact: Garland Land, Director at http://health.state.mo.us/MICA/nojava.html
Center for Health Information Management and Evaluation
Missouri Department of Health and Senior Services
P.O. Box 570
Jefferson City , Mo. 65102
573-751-6272
landg@dhss.mo.us

Strengths and Weaknesses of Approach

Use of the confidence interval may be necessary to report in a web-based data dissemination system when the results in some cells are significantly smaller than in other cells. For example, in a study of mortality related to a specific surgery, some hospitals will have very few surgeries to consider and deaths may be subject to wide variation across time periods. As results are shown some users might think that the institution is doing better or worse than what one would expect given the true underlying population—when the differences in fact may be due to random variation. The confidence interval allows the user to assess whether the rate of mortality for one hospital is a good estimator of the true rate of mortality in the population, allowing greater confidence in the results. For those with scientific backgrounds, confidence intervals (and coefficients of variation) are quite useful in assessing reliability of the outcomes under study.

While confidence intervals offer a good indicator of statistical power, they should generally not be used to draw comparisons across cells because you cannot necessarily interpret the certainty of the statistical significance (Newton and Rudestam, 1999).

The level of understanding of the user of the data must also be a consideration; it may be difficult for the lay public to understand the information provided by confidence intervals.

Use of χ^2 Tests

Most statistical tests of significance are based upon normal distributions and measurements that are in the form of at least interval scales. These conditions are however, not present when there are very small cell sizes or very small populations. There are statistical tests designed to address these conditions—they are called non-parametric or distribution-free statistics. Chi-square tests have several statistical applications. In this context, chi-square test is used to determine the probability of obtaining the observed results by chance. The chi-square test is one of the more frequently used non-parametric tests; it is relatively easy to meet the assumptions for this test. But there are some nuances to be aware of in relation to the number of variables and cell sizes. For example, a simple contingency table which is called a 2 x 2 table would require use of the Continuity Correction rather than a simple chi square test. Also, if the smallest expected frequency[6] for any cell in a 2 x 2 table is less than 5 then one should use the Fisher Exact Test. In a larger table, there is a requirement that no more than 20% of the expected frequencies in the table can be less than 5. Other requirements include: 1) no cells may be less than 1, and 2) no respondent (individual) may be in more than one cell in the table (independence). If the test is invalid due to cell size, then simply aggregate the data to increase the size of the cell. If we affirm that a difference is present in the two samples with a chi square test, then we reject the null hypothesis that the two samples are the same. Thus, if we are examining rates of breast cancer for women in one county versus rates in another county, if the test is significant, we are rejecting the null hypothesis that the two groups (counties) have equal rates of breast cancer in women. Chi square tests are used because rates of breast cancer may be very low in one or both counties, or because one or both counties have a very small population base.

The chi square test allows us to examine differences where the distribution is not normal—a significant result suggests that the difference is not due to error. This means that we can reliably state that the differences between the two populations are not due to chance.

[6] The expected frequencies are calculated for each cell in the table by multiplying the appropriate row and column totals and dividing by N (Foster, 2001).

In public health, statistical tests are frequently used for determining significant trends over time (see State of Washington's Vista/PH system as an example), and for assessing significant differences in rates between groups. The chi square test is a useful tool to assess these differences, even under conditions of small populations or small cell sizes.

There are a variety of different chi square distributions—one that is frequently used in public health is the Mantel-Haenszel Chi Square Statistic. This is used for a stratified analysis of health risks when statistical control of a few variables is required. It can also be useful in exploring more complex relationships and can sometimes be used to effectively quantify risk when there are numerous variables to control. This test can only be used if both variables lie on an ordinal scale. For more detailed information see Kleinbaum, Kupper and Morganstern (1982). An example of how this method is used in public health practice can be found in the following article**, "**Firearm ownership and health care workers," Public Health Reports, May-June, 1996 by Bruce W. Goldberg, Evelyn Whitlock, and Merwyn Greenlick.

Agencies/Systems Using Approach

CDC EPI INFO™ in the analysis section – allows for chi square statistics to be reported in the output. http://www.cdc.gov/epiinfo

State of Washington, VISTA/PH system
http://www.doh.wa.gov/OS/Vista/Statistical_calculations.htm

Pennsylvania Department of Health
Health Statistics - Technical Assistance
(717)783-2548

Strengths and Weaknesses of Approach

While the chi square test is useful in determining whether a significant difference can be found in a contingency table, it still requires at least 5 cases in each cell in a 2 x 2 table, or no more than 20% of cells with < 5 in a larger contingency table. Also, because the chi square distribution is really a family of distributions based on degrees of freedom, it may require further research to assess which distribution (chi square test) is most appropriate for the data under consideration.

Programming to build chi square tests into a system is relatively easy given that it is available in most statistical programming packages, such as SAS and SPSS and other public health oriented statistical packages. However, analysts working with survey sample data may need to use SUDAAN or another package to account for stratified sampling. For more information on this see Brogan (1997) listed in the references section.

Coefficient of Variation

While confidence intervals assist us in understanding the margin of error, it is sometimes not sufficient to assess quality of the estimate. The coefficient of variation (C.V.) is a measure of the stability of the estimate, compared with the magnitude of the estimate.

The coefficient of variation provides a relative measure of data dispersion compared to the mean: Cv = (standard deviation) / (mean) for the normal (bell shaped) distribution. The coefficient of variation has no units. It may be reported as a simple decimal value or it may be reported as a percentage 100 x Cv = (standard deviation) / (mean). Thus, it is defined as the ratio of the standard deviation to its mean. A smaller C.V. suggests smaller relative variability.

For example, if you were looking at performance of two hospitals in terms of number of deaths over a ten year period, you could take the average number of deaths in Hospital A and the average in Hospital B, but because hospital A has a Type I trauma center (more deaths) and Hospital B is a general community hospital (serious cases are transferred out), the confidence interval is not sufficient to determine whether your estimate reflecting performance differences is accurate. Instead, the coefficient of variation provides a relative measure of the data dispersion compared to the mean in both hospitals. Because Hospital B has fewer deaths, they have more instability in an estimate in any given year in terms of the number of deaths (if you look at the data over the 10-year period). Thus, the C.V. provides additional information for an assessment of the reliability of the information.

The Ontario Ministry of Health has set specific respondent numbers for using the Coefficient of Variation in their Ontario Health Survey. The guidelines for the release of their data state that “if the number of sampled respondents is less than 30, the weighted estimate should not be released regardless of the value of the coefficient of variation for this estimate. For weighted estimates based on sample sizes of 30 or more, the coefficient of variation will conclude whether the estimate is unqualified, qualified, confidential or not releasable. Generally, larger sample sizes provide more reliable estimates of health risks and related health behavior.” This suggests that the underlying population must at least contain 30 individuals before using this stability measure. Cell sizes can be less as long as there are at least 30 individuals (cases) in the population.

In conclusion, use of the confidence interval and the coefficient of variation may be necessary to report in a web-based data dissemination system when the system provides data across communities and facilities, and where small cell sizes exist. In addition to a good understanding of how the data will be used, the level of understanding of the user of the data might also be a consideration; it may be difficult for consumers to understand the information provided by confidence intervals.

Agencies/Systems Using Approach

State of Washington, Department of Health
Contact: David Solet (David.Solet@doh.wa.gov)

Ministry of Health, Ontario Health Survey
http://www.cehip.org/DataInfo/

Strengths and Weaknesses of Approach

The use of the coefficient of variation is rarely seen in web-based data dissemination systems, it is less commonly known and used than the confidence interval. It is also more difficult to explain to lay persons using the data.

That being said, it is a useful addition to the confidence interval, and should be considered for inclusion in the web-based system for use by professionals.

E. Software Tools

State and private organizations are developing open source or proprietary software products that apply multiple approaches for improving statistical reliability and reducing disclosure risk of public health data. In addition to the approaches listed above, there are now several new software packages that provide technical support for protecting public health data. The National Center for Health Statistics (NCHS) sponsored the development of disclosure limitation software for two-way tables by OptTek Systems, Inc.. The OptTek software includes the following functionality:

cell suppression

controlled rounding (minimum-distance controlled rounding)

unbiased controlled rounding

controlled rounding subject to subtotal constraints

synthetic substitution (controlled tabular adjustment)

For more information on the NCHS tool contact Larry Cox at NCHS. The second tool was created by RTI International and it is called MASSC^{SM} and it focused on reducing disclosure risk for surveys where sampling methods have been used. For additional information on this tool, contact Dr. Michael Samuhel at samuhel@rti.org .

F. Recommendations

This review of the statistical approaches for both protecting data from disclosure of sensitive information, and increasing the reliability of the data, should be used in tandem with the other two sets of guidelines. It would generally be appropriate to use both approaches (technical/statistical) for reducing disclosure risk. It would also be useful to improve the reliability of the data offered in web-based data dissemination systems.

We suggest that public health agencies review the current methods that are in use in their web-based data dissemination systems and determine whether addition of other approaches would provide that extra protection and result in more reliable information for the user. Yet, we also support the notion of keeping it as simple as possible—while providing the necessary protection.

We also encourage statisticians to expand their efforts on new or enhanced statistical methods to assure that individuals will not be identified via public health web-based data dissemination systems.

In summary, we suggest that a substantial investment will be necessary if public health agencies are going to take advantage of the more advanced statistical methods. Investments could be targeted at upgrades to systems currently in place, additional research on new statistical approaches, or for training state data system developers and data users.

References

Armitage, P. *Statistical Methods in Medical Res*earch, 2e. Boston: Blackwell Scientific Publications, 1987.

Brogan, D.J. Pitfalls of Using Standard Statistical Software Packages for Sample Survey Data. Rollins School of Public Health, Emory University, Atlanta, April 15, 1997

C. Seltiz, M. Yahoda, M. Deutsch and S.W. Cook. *Research Methods in Social Relations*. Holt, Rinehart and Winston, New York, NY, 1967.
Chiang, CL. (1968). Introduction to Stochastic Processes in Biostatistics. New York: John Wiley & Sons.

Chiang, CL. (1961). Standard Error of the Age-Adjusted Death Rate. Vital Statistics Special Reports, 47(9):275-285.

Cox, L.H. (1987) "A constructive procedure for unbiased controlled rounding." *Journal of the American Statistical Association*, 82: 520-524. (lcox@cdc.gov)

Cox, L.H. (2003) "Balancing Data Quality and Confidentiality for Tablular Data". An Invited Paper for the United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians. Luxembourg, April 2003. (http://www.unece.org/stats/documents/2003.04.confidentiality.htm)

Ernst, L.R. (1989) "Further applications of linear programming to sampling problems." Technical Report—Census/SRD/RR-89-05. Washington, D.C., US Census Bureau.

Ghosh, M. and Rao, J.N.K. (1994). "Small area estimation: an appraisal" (with discussion). Statistical Science, 9, pp 65-93.
Gonzalez, J.F., and L.H. Cox. "Software for Tabular Data Protection." Slides dated September 29, 2004. (lcox@cdc.gov)

Jabine, T.B., (1993) "Statistical Disclosure Limitation Practices of United States Statistical Agencies." *Journal of Official Statistics*, Vol 9, No.2, pp 427-454.

Kleinbaum, Kupper and Morganstern. (1982).*Epidemiologic Research* by (Lifetime Learning Publications), Belmont California: Wadsworth, Inc.

Waller, L.A., B.P. Carlin, H. Xia and A.E. Gelfand (1997). "Hierarchical Spatio-Temporal Mapping of Disease Rates." *Journal of the American Statistical Association*, 92 (438:607-617.
Mantel, N. and Haenszel, W. (1959), "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease," *Journal of the National Cancer Institute*, 22, 719 -748.

National Research Council. (2000) "Improving Access to and Confidentiality of Research Data: Report of a Workshop." Committee on National Statistics, Christopher Mackie and Norman Bradburn, Eds., Commission on Behavioral and Social Sciences and Education, Washington, D.C.: National Academy Press.

Newton, R.R., and Rudestem, K.E. (1999). Your Statistical Consultant: Answers to Your Data Analysis Questions. Thousand Oaks, CA: SAGE Publications.

Oakes, M. 1986. *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

R. E. Fay and R. A. Herriott. (1979) "Estimates of income for small places: an application of James-Stein procedures to census data." *Journal of the American Statistical Association*, 74, pp. 269-277.

Shen W, Louis TA (1999). Empirical Bayes Estimation via the Smoothing by Roughening Approach. *J. Computational and Graphical Statistics*, 8: 800-823.

Shen W, Louis TA (2000). Triple-Goal estimates for Disease Mapping. *Statistics in Medicine*, 19: 2295-2308.

Simonoff, J. (1996). Smoothing Methods in Statistics:Springer Series in Statistics. New York: Springer Veriag.

Stoto, Mike. (2003). "Statistical Issues in Interactive Web-based Public Health Data Dissemination Systems.", RAND, presentation at NAPHSIS Meeting, New York, NY, June 2003.

W. Paul Vogt (1993) *Dictionary of Statistics and Methodology: A Nontechnical Guide for the Social Sciences*. SAGE Publications, Newbury Park, NJ.

Zarate, A.O. "The ICDAG Checklist on Disclosure Potential of Proposed Data Releases—A Tool for Disclosure Review." A presentation at the NCHS National Conference on Health Statistics, August 2, 1999.

Management and Institutional Controls for Reducing Disclosure Risk in Web-based Data Dissemination of Public Health Data

Guidelines and Resources for Health Data Organizations

NAHDO-CDC Cooperative Agreement Project
CDC Assessment Initiative

By: Barbara Rudolph, PhD
Luis Paita, PhD.
Gulzar H. Shah, MStat, MSS, Ph.D.

The National Association of
Health Data Organizations

MANAGEMENT AND INSTITUTIONAL CONTROLS FOR REDUCING DISCLOSURE RISK IN WEB-BASED DATA DISSEMINATION OF PUBLIC HEALTH DATA

A. Introduction

Public health agencies are increasingly disseminating health statistics on the Internet. Researchers are increasingly using public health data sets for health services research, including longitudinal and cross-market studies. Broader use and dissemination of public health data sets serves a public good, highlights public health's important role in health and health improvement, and places additional value on public health data assets. Local public health and community based non-profit agencies also rely on easily accessible public health data. Yet, there is more risk of personal disclosure of sensitive information when it is displayed on the Internet; the Internet is an impersonal access tool that increases the velocity of interactions and as a result allows for rapid use and dissemination to others who may or may not have a good understanding of appropriate data use.

What constitutes disclosure? "Disclosure relates to inappropriate attribution of information to a data subject, whether an individual or an organization. Disclosure occurs when a data subject is identified from a released file (identity disclosure), sensitive information about a data subject is revealed through the released file (attribute disclosure)" (Duncan et al., 1993: 23-24)

The confidentiality issues of greatest concern are discovering the identity of someone who is represented in a public health database and discovering that person's personal or medical characteristics through tabulated data. Depending on the nature of the database, the knowledge that someone is in it can itself be harmful. The likelihood of disclosure is higher when there are relatively few people with knowable demographic characteristics such as sex, age, and race in a small community.

This document addressing the management and technical disclosure controls for micro-data in public health web-based data dissemination systems is part of a guideline set—all aimed at assisting the public health data manager in designing or updating web-based information systems. These guideline sets include: Statistical approaches for small numbers, Addressing reliability and disclosure risk in web-based data dissemination, security of data for web-based data dissemination tools, and management and institutional controls for reducing disclosure risk in web-based data dissemination of public health data; as a package. The guideline set will address reduction in the risk of inappropriate disclosure of sensitive information, provision of reliable statistics, and increased security of the data.

B. Summary of Approaches

Within this document there are two broad headings under which a variety of approaches exist. We have labeled the first "management and institutional controls;" the second we titled, “data modification and alteration methods.” Both should be applied to web-based micro-data dissemination systems by public health agencies as they advance and expand their dissemination agendas. In the previous chapter, we describe statistical methods for data modification and appropriate interpretation.

Multiple protective layers to assure anonymity and confidentiality should include the management and technical controls and data modification and alteration described below:

(i) **Data protection agreements:** Data Protection agreements are used to both inform users and control use of the data. In web-based data query systems (WBQS’s) the effectiveness of this approach has not yet been tested.

(ii) **Limited data set:** A subset of the full data set is created for public use, dropping identifiable data elements.

(iii) **On-line query system:** Users are not allowed to download or obtain copies of raw data files. Instead data reside on the host machine often protected by a firewall. The users conduct their own analysis by submitting queries and obtaining aggregated results.
(iv) **User authentication and access validation:** Password protection to CD-ROM and public use files and for access to web query systems.

(v) **Education and training of public use file users:** The data-providing agency educates users about the disclosure risk of micro data, the types of analyses that are considered breeches of confidentiality, and the legal issues associated with disclosure.

(vi) **Making preconstructed tables and pivot tables available.** Pre-constructed tables allow review of results in a form that disclosure of personal identity and health information. Pivot tables, do allow for some alternative displays of the data for the end-user, still controlling the level of drill down to prevent disclosure.(vii) **Anonymizing/de-identifying data**. Anonymizing a micro data file by removing information such as names, addresses, policy numbers, etc.

(viii) **Cross-tabulations and micro-aggregation**: The display is fixed in terms of number of rows and columns and/or the data is aggregated to avoid disclosure.

(x) **Restriction of geographic detail:** Rare events or events occurring in small geographic areas are removed or altered to avoid disclosure.

(xi) **Recoding into intervals and rounding:** Grouping values of continuous data elements into broader categories to increase the cell size and prevent the disclosure.

(xi) Cell suppression: Removing data values below pre-determined cell sizes and applying rules regarding the display of margins.

In addition to the approaches listed above, several new software packages are available that provide technical support for protecting public health data. The National Center for Health Statistics (NCHS) sponsored the development of disclosure limitation software for two-way tables by OptTek Systems, Inc.(OptTek, 2002). The OptTek software includes the following functionality:

- cell suppression
- controlled rounding (minimum-distance controlled rounding)

- unbiased controlled rounding
- controlled rounding subject to subtotal constraints
- synthetic substitution (controlled tabular adjustment)

The second tool was created by RTI International and it is called MASSC^{SM} it focused on reducing disclosure risk for surveys where sampling methods have been used. For additional information on this tool, contact Dr. Michael Samuhel at samuhel@rti.org .

While this document is focused at maintaining control over information, we must remain cognizant of the need for information and therefore, not over-protect the information. As noted in the Introduction, the use and dissemination of public health data serves a public good.

We have artificially separated the statistical approaches and security of the data from the approaches listed above, but the data manager will likely want to incorporate those approaches as well. We have provided references as available for further review, and some examples from public health agencies across the country. We have also described the strengths and weaknesses of each approach.

C. Management and Institutional Controls

There are a variety of tools available for web-based data dissemination which can reduce disclosure risk. These tools vary from user-directed education and agreements to tools that alter the access to data on the system.

Data protection agreements

Public health agencies have been using data release agreements for years; these agreements may be quite restrictive. The Internet has changed the format of data protection agreements (DPA). In open query systems, site users are not required to submit their name and do not need to sign a paper agreement but rather may be required to read a web document and click on a button indicating they have read the document and agree to follow the rules outlined in that document. The effectiveness of this approach has not yet been tested to our knowledge. Not knowing the effectiveness suggests that this DPA is not an approach that could be used as a stand-alone; it must be used in combination with other approaches. In closed web-based systems, data protection agreements are generally required prior to password assignment and log-in.

Agencies Using Approach

South Carolina's Office of Research and Statistics

Massachusetts Department of Public Health

2-Utah Department of Health Indicator Based Information System (IBIS) at http://ibis/health/utah.gov. Contact Lois Haggard, PhD:loishaggard@utah.gov

Strengths and Limitations of Approach

The DPA establishes institutional control, is compliant with HIPAA Privacy provisions for a limited data set, and serves to educate public health staff and data users as to their due diligence and legal obligations to protect the data and properly use the data.

Limited Data Sets

Limited datasets based on reduction in the number and type of data elements is a likely choice for most web-based systems. In creating the limited data sets, certain variables are either not included or are modified. For example, ages beyond 65 are aggregated or age groups rather than single-year ages are reported. While this can provide a reduction in risk it also limits the types of questions that can be answered from the remaining data. It is one of the methods suggested by HIPAA regulations for the release of health data by covered entities. Many public health agencies are exempt from this regulation; however, it is likely that most public health agencies are influenced by it, or operate under public health laws with similar requirements.

References

HIPAA Privacy Rule http://www.hhs.gov/ocr/hipaa/

Agencies Using Approach

AHRQ's HCUP system for hospital discharge data http://www.ahrq.gov/data/hcup/

Numerous state agencies

Strengths and Limitations of Approach

A set back from this approach is the loss of specificity in some data elements. Further, due to elimination of confidential data elements, for linking of data across time or institution are not available for longitudinal studies, or for episode of care analyses. An advantage is that a limited data set streamlines the data acquisition process by not requiring IRB approval and yet still supports most statistical studies.

On-Line Query Systems Limits

Web systems use data modification and alteration methods and rely on limited datasets to ensure protection of native files. It is very important for system developers to create a new dataset that is separately housed from the native file to prevent file corruption and access by unauthorized users. Query systems can also be designed to return limited tables and pivot tables, also limiting the risk of identification. Systems can also be configured to limit access to micro-data through logon access and upfront data user agreements.

Agencies Using Approach

Florida: http://www.floridacharts.com/charts/chart.aspx

Kansas: kic.kdhe.state.ks.us/kic/

Missouri: http://www.health.state.mo.us/GLRequest/MICAdef.html

Pennsylvania: http://www.phc4.org/Default.htm

Tennessee: http://oit.utk.edu/helpdesk/

Utah: http://www. http://ibis.health.utah.gov/

Washington: Vista/PHw - Washington State Center for Health Statistics (www.doh.wa.gov/OS/Vista/homepage.htm)

https://fortress.wa.gov/doh/epiqms/

Wisconsin: http://www.dhfs.state.wi.us/healthcareinfo/qsmain.htm

Strengths and Limitations of Approach

A clear strength is the system's ability to help the novice user query the database and generate custom-made tables "on the fly" in seconds, without requiring any programming or statistical skills on the part of the user. Suppression of numbers for rural areas and subgroup characteristics results in loss of information to user. Queries may not support detailed analyses, but rather serve as a preliminary study tool.

User Authentication and Access Validation

It is possible to implement password protection to CD-ROM and public use files and for access to web query systems. Other less restrictive alternatives include simply requiring registration of the user for each use. The least restrictive option is to limit the access to the system to only those who have completed or agreed to a data use agreement. For instance users of IBIS see a pop-up window containing a data use agreement. The users are only required to click o the button marked "agree".

Agencies Using Approach

Missouri Department of Health and Senior Services

Missouri Information for Community Assessment, (MICA)
Contact: Garland Land, Director
Center for Health Information Management and Evaluation
Missouri Department of Health and Senior Services
P.O. Box 570
Jefferson City , Mo. 65102
573-751-6272

Utah Department of Health, Indicator Based Information System (IBIS)
Contact: Lois Haggard, Ph.D., Utah Department of Health
loishaggard@utah.gov

Strengths and Limitations of Approach

This is relatively inexpensive and technically it is not difficult to implement. This technique provides added protection to data access and supports tracking of users. Administration costs include set up and maintenance of a logon/identification process and potentially monitoring of use. A limitation is that by allowing access just click of "agree" button, the access is not restricted in practice. The users may not even read the agreement.

Education and Training of Public Use File Users

Some web-based systems are complex enough to recommend that there is appropriate training of users. This will limit the number of users of the system—unless the training mechanism is also a web-based system. For example, the user could be required to pass a short test taken from training material on the website. This is a technique used by a number of large universities in regard to human subjects' provisions. Many Federal agencies have "user training" for database users. Medicare has established training centers for users of Medicare claims data. New methods such as web based training sessions can be done using technology for "WEBINARS". Training may also be used to serve an important latent function- -sensitizing users to importance of confidentiality and prevention of exposure and penalties associated with data misuse.

Agencies Using Approach

Washington State Department of Health, Data Users Conferences

National Center for Health Statistics, NHIS data file user training

AHRQ, HCUP users training http://www.ahrq.gov/data/hcup/

Strengths and Limitations of Approach

In-person training requires a substantial investment by the data providing agency. Less costly methods can be implemented via web seminars. Those who attend training may get better access to data with less modifications/alterations, thereby permitting improved analyses. Training requirements also place limitations on the number of individuals who can be certified to use the data, especially if in-person methods are used. Few individuals can invest the time and expense to travel to attend training sessions.

Pre-constructed Tables and Pivot Tables

Some query systems are constructed to produce only those tables that have been pre-designed by the data agency. Others allow the user to implement the pivot functionality.

Strengths and Limitations of Approach

Both of these approaches limit the types of queries and the output from those queries, protecting the data from misuse. These forms of output would require that a significant number of queries would be run before one could potentially put together all the underlying data on an individual within the data, and to learn something new about that individual or to be able to identify the individual within the data.

D. Data Modification and Alteration

These are various technical strategies to protect public health data and include methods that can modify or alter the data file, reducing the probability that individuals can be uniquely identified in some way.

Data modification and alteration methods can be relatively complex, although the most complex methods are usually associated with the need for better statistical reliability (this is discussed further in the "Statistical Approaches for Small Numbers: Addressing Reliability and Disclosure Risk" document. Their application to public health data sets may require significant new programming of many web-based systems and analytic/source file programs. These techniques will significantly alter the information available in the individual micro-level records and could reduce the utility of the data sets to some of their primary customers. This, however, may be a better option compared with alternatives such as data aggregation and suppression of cells containing small numbers. While it may be possible to provide more details given the use of these new methods, it may come at the cost of statistical versus "real" data.   Each method has strengths and weaknesses (See Strengths and Limitations of Approach).

Described below are methods which are based only on alterations to the existing data—not included are techniques that swap in data from other geographic locations or synthetic data from statistical modeling approaches—these approaches are found in the "Statistical Approaches for Small Numbers.." document.

Anonymizing/de-identifying data: Anonymizing a micro data file by individuals is the most common method of data modification.

Cross-tabulations and micro-aggregation: Data are presented in tabular format (individual data are not released). For continuous variables in the data, means, variances, and covariances may be released.

Restricting geographic detail: For rare events resulting in small numbers, geographic details may not be made available.

Recoding into intervals and rounding: Grouping values of data elements that are continuous (e.g., date of birth recoded into age categories), resulting in ordinal variables with discrete values.

Cell suppression: Removing data values from the cell based on pre-determined cells sizes and rules regarding display of margin.

References

Risk of Disclosing Individually-Identifiable Information from Public Use Hospital Patient Discharge Data Files, Braday. H., Duffy, L., Powell, A., UC Data Archive and Technical Assistance, UC Berkeley, March 2002.

Discharge Data: Assuring Confidentiality While Providing Timely and Meaningful Information—is it possible? Rudolph, B., University of Wisconsin, 2003.

Anonymizing/de-identifying Data Files

State agencies often use various encryption algorithms for unique patient identifiers, transforming the identifier into a stable unique number. This number and its' linkage to the individual are separated and stored apart from each other in locked files, bank vaults, or other secure arrangements. This is also a requirement of HIPAA privacy regulations. For example, one state agency in New York created unique identifier by combining last few digits of social security number (SSN) with first two letters from first name, two letters from last name, and parts of date of birth.

References

UC Data Archive & Technical Assistance, February 2002. http://odwin.ucsd.edu/idata/

Agencies Using Approach

At least 17 state health data agencies use an encrypted ID.

AHRQ HCUP data system http://www.ahrq.gov/data/hcup/

Strengths and Limitations of Approach

This method by itself may not adequately control disclosure risk since other characteristics included on the file could be used to associate or "construct" an individual's identity with a record on a micro data file. For example, probabilistic matching techniques using variable such as age, gender, Zip code, date of hospitalization discharge, etc., can be used to link individuals to events when there are public records of the event (motor vehicle accident, other highly unusual circumstances), or knowledge of the individual. The benefits associated with an encrypted ID include being able to link across healthcare events allowing alternative forms of analysis such as analysis of episodes of care for chronic conditions.

Cross-tabulations and Micro-aggregation

Nearly all systems produce aggregated cross-tabulations in order to reduce the risk of identity. For example, individuals can be aggregated according to age groups and gender categories for each of the cross-tab columns (or rows) depending on the question asked. This makes it very difficult to identify the individuals within those cells, as long as there is a large enough population and adequate cell sizes. Depending on the size of the underlying population and the statistic being used, what makes an adequate cell size can range from 3 to 30 cases.

Agencies Using this Approach

Nearly all state data organizations use aggregations to address small cells and data reliability in their web-based systems.

Strengths and Limitations of Approach

Aggregation creates an acceptable cell size for statistical purposes. However, in order to achieve this result, specific information may be lost in the aggregation process. Aggregation should be carefully applied given the projected specific purposes for the web-based data system. It can result in serious loss of information for analysis, yet identity could be inferred from tables by using multiple tables if no other techniques are applied.

Restricting Geographic Detail

An example of this type of restriction of geographic data elements is when in-state zip and out-of-state zip codes with less than 30 discharges in a calendar year are coded at the county or state level respectively. This reduces the probability of identifying an individual based on their location within a small zip code. This type of reduction in information can be problematic however, for those seeking information on rural areas and out-community level.

Agencies using this Approach

Utah Office of Healthcare Statistics, Utah Department of Health, Center for Health Data, Utah Department of Health, Salt Lake City, UT 84114-2101, Phone: 801-538-9191 or contact: loishaggard@utah.gov

Wisconsin Bureau of Health Information, Department of Health and Family Services http://www.dhfs.state.wi.us/healthcareinfo/qsmain.htm

Strengths and Limitations of Approach

This approach creates a greater pool of individuals within a geo-area, allowing statistical tests to be used. Community level information may not be available for planning at that level.

Limiting the Number of Data Elements in a Micro File

For example, this may be used for special handling of sensitive diagnoses: age, sex, and zip code are encrypted if the discharge involves Major Diagnosis Code (MDC) "25-Human Immunodeficiency Virus Infection" or Diagnosis Related Groups (DRG) "433, 521-523 - Alcohol/Drug Abuse or Dependence." This assures that individuals with HIV cannot be identified from use of the file.

Strengths and Limitations of Approach

This limits the probability of uniquely identifying an individual, while preserving useful information for health assessment, health planning, and utilization studies. There may be a loss of specificity for some research and public health applications.

Recoding into Intervals and Rounding

In this approach, continuous variables are re-coded into categorical variables with few attributes (categories). For example, date of birth may be mapped into 5 year age categories; or individuals over 80 years of age are grouped together, while younger ages may be in 5 year categories. This is done to prevent disclosure of individuals in categories where there are only a few individuals. Rounding might be used for age, or for variables such as family income.

Agencies Using Approach

Numerous state agencies including:

Utah Office of Healthcare Statistics, Utah Department of Health, Center for Health Data, Utah Department of Health, Salt Lake City, UT 84114-2101, Phone: 801-538-9191 or contact: loishaggard@utah.gov

Wisconsin Bureau of Health Information, Department of Health and Family Services, http://dhfs.wisconsin.gov/stats/queries.htm

Strengths and Limitations of Approach

Many state agencies recode dates (birth, admission, discharge) for web-based data dissemination systems, adding a protective layer to the data by reducing risk of re-identification. However, researchers may need exact dates for linking or specific analyses, or the submitting healthcare provider may need information for quality assurance activities. With proper permission and encryption codes, the identifiable information can generally be reconstructed for linkages, etc.

Cell Suppression Methods

De-identified health information displayed in tables, whether web-based or document-based, can still result in re-identification when cell sizes are small. The primary means for protecting confidentiality in web-based data dissemination systems, as in more traditional dissemination systems, is the suppression of "small" cells, plus complementary cells, in tables. This approach often results in a substantial loss of information and utility. Alternative approaches include "perturbation" methods such as "data swapping" and "controlled rounding" that can limit disclosure risk while maximizing information available to the user. These approaches are described in the "Statistical Approaches for Small Numbers: Addressing Reliability and Disclosure Risk" document.

E. Suppression Rules

In this section, we will describe several approaches used in public health agencies for suppression algorithms, the reader should note that the Missouri/Garland Land approach has been supported by the National Center for Health Statistics. Arguably, each approach "rule" has both flaws and strengths.

"The Numerator Rule"

The numerator rule is designed to prevent the release of information when there are fewer than *x* individuals in a given category to be used as a numerator in the calculation of rates and ratios. Complementary categories (cells in the same row or column of a small cell) must also be suppressed to avoid discovery of the number of cases by subtraction from marginals (cell totals). For instance, suppose there were 10 AIDS deaths among men in a small community. Reporting that 9 of the decedents were White men is tantamount to saying that 1 was Black. With complementary suppression data quickly become unusable.

The best rationale for "*numerator-based*" data suppression is confidentiality protection, not statistical reliability. Suppression rules generally work well in protecting identity but may not prevent someone trying to uncover certain characteristics. Because marginal counts or complementary cells, including ones with large numbers, must also be suppressed in order to prevent calculation of the non-reported cell, the information lost can be substantial. (See Stoto, 2002:32). There are algorithms to minimize the number of complementary cells that must be suppressed, but they do not guarantee non-identifiably (Federal Committee on Statistical Methodology, 1994).

Agencies/Systems Using Approach

Utah Department of Health Indicator-Based Information System (IBIS).
(Threshold N=5)
Contact: Lois Haggard, Utah Department of Health
loishaggard@utah.gov

Vista/PHw - Washington State Center for Health Statistics
(Threshold N = 5)
Contact: David Solet, Washington Center for Health Statistics.
(www.doh.wa.gov/OS/Vista/homepage.htm)

VitalNet
(User-specified threshold)
Contact: Daniel Goldman, System Developer, Expert Health Data Programming
(www.ehdp.com/vitalnet/)

“Numerator-based Cell Suppression Variations”

In this variation of the “*numerator-based suppression*” rule, not only are all statistical cells with one to five subjects suppressed, but there is additional suppression of all statistical cells that would allow for the calculation of any other cells with values of 1-4 (Cohen, 2001). While the suppression helps in protecting individual identity it also results in loss of information.

Agencies/Systems Using Approach

Massachusetts Department of Public Health
See Confidentiality Policy and Procedures

The “Denominator Rule”

The *“Denominator Rule”* is designed to prevent the display of information when the population under consideration is less than a certain size, such as 100,000 population. The assumption made is that there are a limited number of persons with any given set of characteristics in a small population, therefore by extending the population covered to a larger size, one can protect the identity of individuals. Overtime a number of state agencies have used as a minimum population 30 cases/events. When the denominator is less than 30, the cell is suppressed. No attention is then paid to the actual cell size. This can pose problems for statistical testing—the reliability of the result may be questioned.

Additionally, confusion may exist about what DENOMINATOR means:

Is it the number in the POPULATION?; or

Is it the number of EVENTS (e.g. deaths of any cause)?; or

Is it the number if deaths of any cause in a certain age group or geographic area?

One disadvantage of using the denominator rule alone (as opposed to in combination with the numerator size rule) is that there’s a potential for a table with adequate numbers in all its cells to be suppressed. One possible solution is to restrict the application of this rule to rare events (which means also small numbers in the numerator) or extremely skewed distributions.

Agencies/Systems Using Approach

Indicator-Based Information System (IBIS).
(health.utah.gov/ibis-ph)
Utah surveys using this method:

YRBS: National, unweighted denominator < 50 cases then:
--include 95% confidence interval when reporting percentage/means
--suppress estimate and footnote (estimate based on <50 cases and is unstable)
--group variable to increase number (e.g. combine grades)
State/Local: Estimate suppressed when denominator < 50 cases
Contact: Lois Haggard, Utah Department of Health
loishaggard@utah.gov

"Numerator and Event Denominator Rule"

This rule allows the display of only the marginal values of a table if the difference of the value of any table cell and the corresponding number of total events in the data file for individuals with the same characteristics is a small number , e.g., less than 10 (Land, 2001)

For example, a cell with one AIDS death of an African American female aged 25-34 would be published if there were 15 of African American female aged 25-34 total deaths. The assumption is that it may be possible to identify the diagnosis of a person if there are fewer than 10 people with the same demographics characteristics and who had the same event (death, in this case, or perhaps birth or hospitalization).

In addition, if less than two row or column totals are greater than five then all the row or column totals are suppressed. This additional rule prohibits determining a suppressed table if the margin totals are small.

This rule protects against release of data when there is a small difference between the number of events in a cell of a table and the total events related to the cell. The rule however, does allow for some small numbers to be displayed when there is a large difference between the events displayed and the total events related to the cell.

This rule requires an interactive determination of the total cell counts of the file to compare with the proposed table. This is a major disadvantage if one is attempting to build an open query systems. Algorithms might be designed to address this, but the cost of doing so would be high.

Agencies/Systems Using Approach

Missouri Information for Community Assessment (MICA).
Missouri Department of Health and Senior Services
Contact person: Garland Land. Director
Center for Health Information Management and Evaluation
Missouri Department of Health and Senior Services
P.O. Box 570
Jefferson City , Mo. 65102
573-751-6272
landg@dhss.mo.us
(www.dhss.mo.gov/MICA/nojava.html)

Criteria Used in Utah:
≥ 100 persons are in population of interest
≥ 20 cases in the numerator are in population of interest

Utah Department of Health
Contact: Lois Haggard, Utah Department of Health
loishaggard@utah.gov

Numerator/Denominator-Based Suppression

Cell sizes based on a combination of denominator (population from which the health events arise) and numerator (health event) are suppressed in accordance with the table shown below. Aggregate data with denominator and numerator values greater than those indicated in the table may be considered sufficiently de-identified so as not to constitute confidential information, and may be disclosed(Cohen, 2001). This method has a similar weakness in regard to the need to apply the method and then determine whether there is any additional privacy risk, this precludes interactive non-restricted query systems. These systems, however, could likely be designed with restricted display of output, and other technical measures (micro-aggregation) to control for release of additional information about the individual.

DENOMINATOR (D)	NUMERATOR (N)	STANDARD
10-29	1-4	Suppress numerator and any other cells[6] that would allow for the calculation of any other cells with values of 1-4
10-29	5-29	Suppress any cells that would allow for the calculation of any other cells[6] with values of 1-4
0-8	0-9	Suppress numerator
=N	=D	Suppress numerator unless privacy risk is minimal

Agencies/Systems Using Approach

Massachusetts Department of Public Health
See Confidentiality Policy and Procedures

Alternative Suppression Standards

In the Missouri Department of Health and Senior Services, any Center may develop an alternative aggregate data release standard if it decides not to follow any of the standards above, provided that the standard is at least as restrictive as the above stated standards (see "*Numerator and Event Denominator Rule*" Missouri); and any alternative standard is documented by the Center and approved by the Privacy Officer prior to implementation. This approach provides flexibility for special circumstances and merged databases.

In many public health agencies, suppression standards are based on the specific database, mandates via funding organizations, history of the data release, preferences of specific data stewards. This can cause problems when databases are merged to answer specific questions, for example, if a cancer registry has a "denominator" rule of 1000 and a hospital discharge system has a "numerator" suppression rule of <5 in a cell, both databases could "charge inappropriate release of information" when a merged file is created for web-based data dissemination. The solution is as stated above, prior approval by a Privacy Officer who can mediate the two alternative rules.

Agencies/Systems Using Approach

Missouri Department of Health and Senior Services
Contact: Garland Land Director
Center for Health Information Management and Evaluation
Missouri Department of Health and Senior Services
P.O. Box 570
Jefferson City , Mo. 65102
573-751-6272
landg@dhss.mo.us

F. Summary

This guideline addresses the various options for reducing disclosure risk for public health data in web-based data dissemination systems. The combination of methods is up to the user given the context of their environment, data system constituents and mandates, type of user web access, and assessment of risk of disclosure. In the attached Appendix there is a decision-tool for assessing risk of disclosure.

References

Cohen, BB. (2001). Guidelines for the release of aggregate statistical data: Massachusetts perspective on issues and options. Presentation at the Assessment Initiative/NAPHSIS Conference, September 12, 2001.

Doyle P, Lane JI, Theeuwes JJM, and Zayatz LM, eds., (2001). Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies. Amsterdam: Elsevier Science BV.

Duffy, Braday. H., L. (2002) Powell, A Risk of Disclosing Individually-Identifiable Information from Public Use Hospital Patient Discharge Data Files, UC Data Archive and Technical Assistance, UC Berkeley.

Duncan GT, (2001). Confidentiality and statistical disclosure limitation. In *International Encylopedia of the Social and Behavioral Sciences* (cited in Duncan et al., 2001).

Federal Committee on Statistical Methodology, (1994). Statistical Policy Working Paper 22 – Report on Statistical Disclosure Limitation Methodology. Washington: Statistical Policy Office, Office of Management and Budget.

Fienberg SE, Makov UE, Steele RJ, (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics* 14: 485-502.

HIPAA Privacy Rule online at: http://www.hhs.gov/ocr/hipaa/

Land, G. Dec. (2001). *Confidential data release rules.* Presentation to WDDS Users Network.

OptTec. (2002). OptTek Systems, Inc. Receives Contract Award to Develop and Test Implementation Software For Confidentiality Protection: Press release. Online, retrieved Nov 07, 2005 from http://www.opttek.com/news/index3.html

Rudolph, B. (2003). Discharge Data: Assuring Confidentiality While Providing Timely and Meaningful Information—is it possible? University of Wisconsin, 2003.

Stoto, Michael. (2002). Statistical Issues in Interactive Web-based Public Health Data Dissemination Systems, RAND Health, September 19, 2002.

UC Data Archive & Technical Assistance, February 2002. http://odwin.ucsd.edu/idata/

Wilson, David. (2004). "Protecting Quality and Confidentiality of Data by MASSC- - A Survey Sampling-Based Method." Presented at RTI 2004 Joint Statistical Meeting (JSM) Conference. August 9, 2004, Toronto, Canada.

A Guide to Designating Geographic Areas for Small Area Analysis in Public Health: Using Utah's Example

Guidelines and Resources for Health Data Organizations

NAHDO-CDC Cooperative Agreement Project
CDC Assessment Initiative

April 2005

By: Gulzar H. Shah, MStat, MSS, Ph.D.

THE NATIONAL ASSOCIATION OF
HEALTH DATA ORGANIZATIONS

NAHDO

A GUIDE TO DESIGNATING GEOGRAPHIC AREAS FOR SMALL AREA ANALYSIS IN PUBLIC HEALTH: USING UTAH'S EXAMPLE

A. Introduction

This document offers a guideline for states to develop a small areas scheme for analysis of health data, using small area analysis framework. The author primarily draws from his experience of developing small area analysis at the Utah Department of health (see Haggard, Shah and Rolfs, 1999)[7]. The term 'small area' is used to imply areas that are large enough to have a sufficient number of events of interest to yield stable rates, yet they are small enough to unmask variations in the rates and still convey a sense of community.

Public health policy has increasingly emphasized local, or community health assessment and planning (American Public Health Association, 1991; APEXPH Steering Committee, 1991; Stano, 1993). These efforts are often hampered by a dearth of relevant and meaningful information about the current health status and needs of local populations. Understanding community health status at the small area level can help policy makers improve community public health planning. Several functions of small area analysis render analyses at this level useful at various levels.

Small area analysis has emerged as a useful tool in health services research over the last two or three decades, however, the history of its use is more extensive (see e.g. Glover 1938, as cited in Goodman and Green, 1996) It is a useful tool to describe how rates of health care use and events vary over meaningfully defined geographic areas. The tool has been used to investigate variation in the rates of hospitalization for a large array of diseases and surgical procedures including: chronic obstructive lung disease, pneumonia, hypertension, and in surgical procedures, such as hysterectomy, cholecystectomy, and tonsillectomy. Among the potential sources of geographic variation are differences in underlying morbidity, access to care, physician judgment, quality of care delivered, patient demand for services, differences in the supply of medical care resources, such as hospital beds, and uncertainty in the outcomes of different diagnostic and therapeutic procedures (Parchman, 1995; Goodman and Green, 1996). Such use of small area analysis, can lead to improved medical care (Goodman and Green, 1996; Kazandjian and Hudson 1990).

A variety of methods are used in creating a small area analysis scheme in aggregating an appropriate number of persons into discrete geographic units below the state level. As stated by Haining, Wise and Blake (1994), in constructing small areas for the analysis of health data, the small area framework should enable the analyst to link health data and census data. Further, the areas should have large enough populations to ensure that rates are reliable and be homogeneous with respect to important socio-economic attributes.

[7] The author is indebted to Dr. Lois Haggard, Director Office of Public Health Data, Utah State Department of Health and Dr. Robert Rolfs, State Epidemiologist, Utah State Department of Health, for their contribution and guidance in the process of development of Utah small area analysis scheme.

B. Steps in Developing Small Areas

Developing a small area analysis scheme involves several steps, and each state may face different set of challenges in developing their own. The challenges will vary partially due to the **individual state's resource base** and **analytical capacity**, and partially due to availability of data and information necessary for these analyses.

The procedures outlined in this paper may be implemented with considerable variations. For instance, in Utah, race/ethnicity was not an important variable because most of Utah's population is Caucasian. There are no significant ethnic enclaves in any of the geographic areas. However, in some other states where race/ethnicity based subgroups are concentrated in reasonably large sizes, considering race/ethnicity as a criterion for grouping small areas may have been imperative. In sum, variations in individual states' methodologies are likely and even expected; this paper attempts to provide solutions based on Utah's experience.

Obtain the Population Estimates for all Administrative Boundaries, Understand Your Need:

The first step in the process of deciding small area boundaries is to understand the existing administrative boundaries. For this, population estimates for the most recent years are required for various levels of aggregation, such as five-digit zip codes, county, and districts. In theory, a state may already have an evenly distributed population across existing administrative boundaries such as counties. Alternatively, the urban zip codes may be roughly equal in size and may have a large enough population to serve as independent small areas; rural counties may be of appropriate size to be considered candidates for a small area analysis (see for instance, Joines et al., 2003). To illustrate why Utah felt a need for small area analysis, Utah's situation is outlined.

Utah's problem and need: Utah has 12 local health districts and 29 counties. It could have been convenient to use local health districts or counties as aggregation unit in the analysis of health data, such as hospital discharge data, the Medicaid data, and vital statistics, but there were problems associated with such aggregation. Both local health districts and counties are characterized by extreme stratification with respect to their size and density. While urban counties – Utah, Weber, Davis, Salt Lake -- have a population density of 424 persons per square mile, frontier counties have only 2.5 persons per square mile. The rural counties also lack population density (15.3 persons/ sq. mile). As far as relative size is concerned, the four urban counties contain nearly 77% of the entire state's population. The 15 frontier counties, on the other hand contain as little as 7% of the population with the remaining 16% of people residing in the 10 rural counties. The Salt Lake City County alone contains 41% of the population. Computing and comparing rates for many health indicators and rare events was statistically in appropriate in that the larger counties would mask community level disparities on one hand, and the smaller county would have unstable rates on the other.

The Nature of Condition, Disease, or Event for Which Small Area Analyses are to be Developed.

Although carefully developed areas for infrequent events can be used for relatively more frequent events as well, the nature of the small areas may vary depending upon the type of public health events of interest, number of years for which data are available (schwarts et al., 1994) and the geographic variable available in the dataset containing the event of interest (see point 6.1 later on in this paper). Identify what geographic boundaries are available in the public health data for defining the small areas. For instance population estimates are available for census blocks and census tracts. However, neither block groups nor census tracts can be used in most states because health data such as vital records, hospital data, and health surveys are not identified by census block group or tract.

In Utah, zip code areas were used to define small areas because they are the smallest commonly-used geographic units that are also identified in most health data sources. Zip code areas are discrete geographic areas used by the U.S. Postal Service in mail delivery that often roughly follow administrative boundaries. In some sparsely populated areas, counties were used as the geographic unit of interest.

Obtain the Maps

As stated earlier, public health data for most of the states contain zip code, city or town, and county information.

The next step is to decide if zip code can be used as the building block of small areas in your state.

If so, obtain maps of zip codes and corresponding neighborhood names. The zip code maps, along with corresponding boundaries for neighborhood (municipality, city/town council etc.), county, and districts etc. are generally available from the post office.

Obtain the history of relatively newer zip codes to determine when they were developed. Not knowing the time of new Zip codes development can be a source of confusion in matching the zip code characteristics (stay cool; this is discussed later) to the public health data records, with many orphan codes[8] in the former. Also, there are zip codes that grow large enough that they split into two or sometimes multiple zip codes over time. Obtaining such information is also important.

For **Utah,** the zip code information was readily available through a published document of the postal office. Since such publications are not always available for the latest year, as was true in Utah. The updates were obtained by the researchers through repeated phone calls.

[8] The orphan codes in this context are zip codes that are in the data set but are not in the list of valid zip codes obtained from the U.S. Postal Services.

Obtain Population Characteristics for Five-digit ZIP Codes Code Level:

The next step is to obtain zip code level population characteristics that will be used as criteria for defining small areas. These characteristics include, but are not limited to, any background characteristics of the area, such as age and gender specific (single-year age) population size for **multiple years**, average (median) household income, measure(s) of education level, median age, income, household size. The zip code level information is needed for two purposes:

Small area development: For deciding which zip codes are similar on certain characteristics, to qualify for combining and making a small area;

Small Area Analysis: There are many zip code level characteristics that can be used in the data analysis of the small areas.

Zip code level population estimates and population characteristics are available from the U.S. Population Census. These zip code level characteristics are presented in Table 1 below:

Table 1. Characteristics of the population available at zip code level from the U.S. Census

Population Characteristics	**Measures**
General Demographic Characteristics	total population, gender, median age and other age groups, race, household size, family size, percentage of households occupied by the owners as opposed to renters,
Special Characteristics	Population 25 years and over High school graduate or higher Bachelor's degree or higher Civilian veterans (civilian population 18 years and over) Disability status (population 21 to 64 years) Foreign born Now married (population 15 years and over) Speak a language other than English at home (5 years and over)
Economic Characteristics:	In labor force (population 16 years and over) Mean travel time to work in minutes (population 16 years and older) Median household income (dollars) Median family income (dollars) Per capita income (dollars) Families below poverty level Individuals below poverty level
Housing Characteristics	Single-family owner-occupied homes Median value (dollars) Median of selected monthly owner costs With a mortgage Not mortgaged

Note: The U.S. Census Bureau, Summary File 1 (SF 1) and Summary File 3 (SF 3) contain the Zip Code level information. Information on individual zip codes is available through the American GactFinder at http://factfinder.census.gov/home/saff/main.html?_lang=en

If population estimates are not readily available from the U.S.Census for the required year, there are private vendors who create such data for GIS and other zip code level analyses.

Utah's case: At the time **Utah** developed their small area scheme, U.S. Census did not have population charracteristics and estimates at the zip[9] code level. The Bureau of Surveillance and Analysis, Utah Department of Health purchased their data on population size, median age, and median income for then current Utah zip codes from a commercial vendor, CACI[10] Marketing Inc. CACI constructed population estimates at the zip code level by using the most recent decennial census data and additional information, such as sub-county estimates of change from the U.S. Census Bureau, special censuses, local sources of information about change, and changes in residential delivery statistics from the U.S. Postal Service. Estimates included 1997 population totals and population by sex and age group for each zip code, allowing for age standardization. The CACI file also included estimates for the average annual rate of population change for each zip code area, which allowed for the derivation of 1994 through 1996 population estimates required for the analyses.

Create a Workgroup:

Creation of small area analysis involved some critical decision making and at this stage having a workgroup or an advisory committee can be beneficial. The committee should consist of people who not only understand public health data but also have knowledge about the geographic areas.

Decide about the Population Size for Small Areas:

The next steps are to decide on the criteria for small areas, preferably by convening the workgroup meeting, or through solo effort later to be shared with the workgroup, if a workgroup is in place. The attributes of small areas requiring critical decisions are:

Population Size: What should be the rough population size for a typical small area, given the purpose of analysis?

Other attributes of the zip code: What other attributes of zip codes should be used when deciding about which ones to combine together? These attributes could be sharing a geographic boundary (being adjacent), being part of a certain neighborhood, having similar socioeconomic situation, or some demographic characteristics.

[9] In future, it is worthwhile for health data agency to consider the use of ZIP Code Tabulation Areas (ZCTAs™), recently developed by the U.S. Census Bureau. The ZCTAs are a new statistical entity developed for tabulating summary statistics from Census 2000. ZCTAs are generalized area representations of U.S. Postal Service (USPS) ZIP Code service areas. Each one is built by aggregating the Census 2000 blocks, whose addresses use a given ZIP Code, into a ZCTA which gets that ZIP Code assigned as its ZCTA code. They represent the majority USPS five-digit ZIP Code found in a given area. For more information, please refer to the http://www.census.gov/geo/ZCTA/zcta.html

[10] **CACI International Inc,** David Huffman, Managing Director, CACI Marketing Systems Group, (800) 292-2224; **dhuffman@caci.com**; **1100 North Glebe Road; Arlington, VA 22201; 703-841-7800**
Fax 703-841-7882

The **population size** is generally decided based on the typical events of interest, which will become numerators in the analysis. It is appropriate to use less frequent (even rare) events for this. For instance, for birth certificate data, use of infant deaths or maternal deaths is recommended as a criterion rather than a more frequent event such as prenatal care in the first trimester. The idea is to have the small areas large enough to contain a sufficient count at least annually for the most events of interest collected through public health data sets. Smaller areas may be more meaningful to communities, but rates based on small numerators are unstable (Buescher, 1997) and confidence intervals for such rates are large, rendering the comparisons difficult to interpret for most practical purposes. Using such small areas with small numbers of events may also pose privacy problems for more sensitive events, such as suicide or AIDS. A **numerator of 20 or greater** produces relatively stable estimates, and also approximates a normal distribution of the Poisson parameter (μ), which simplifies computation of confidence intervals (Kahn and Sempos, 1989).

Population size criteria for designing the **Utah small areas** were determined based on health event incidence rates. The population size criteria were determined by examining the three- and five-year incidences of selected events, such as infant mortality and lung cancer, for which small area estimates were desired. It was determined that areas with 40,000 to 60,000 persons would produce incidence counts of 20 or more for a wide range of health events. Increasing the population sizes sufficiently to produce reliable estimates for rare events (e.g., homicide or AIDS) would increase area size beyond that which would allow meaningful community level analyses. Where possible, areas with 40,000 to 60,000 persons were established, but areas with population sizes of approximately 20,000 were created when low population density, community identity, or others factors suggested that it was appropriate.

Select the Large Enough Zip Codes as Small Areas:

Generally, some zip codes in urban areas are large enough to be used as stand alone small areas. For instance, in Utah, 14 of the 61 small areas were stand alone zip codes. Designate the zip codes meeting the minimum population criteria as independent small areas. Then, evaluate the remainder of the zip codes for combining.

Combine Zip Codes with Population Smaller than the Minimum Desired for a Small Area:

Several criteria can be used in a certain pre-decided priority to combine the zip codes that do not qualify for being stand alone small areas. The criterions are (a) political boundaries, (b) being adjacent or being part of a certain neighborhood, (c) having similar socioeconomic situation assessed using a measure like median household income, and (d) some demographic characteristics such as median age, education level etc. While any of these four criteria can be used in a decided priority, the discussion here follows the principles used in Utah.

Administrative boundaries are important criteria to combine zip codes for small areas. For instance, it may not make sense to combine zip codes cutting across local health districts or in some cases even the counties. At this stage decide what administrative/political boundaries the small areas should not cut across. This can be used as the most imperative criterion in combining the zip codes. If certain other characteristics of the zip codes make more sense in your context use that instead. However, within a geographic area, we need more than just one criterion. Before we discuss those in the next section. For 'political boundary' as a criterion, Utah's example below will offer guidance.

In developing the **Utah small area** scheme, some **Utah** counties were found to be too large to yield meaningful rates. Some rural counties were too small to have an aggregate population sufficient for producing stable rates. The zip codes had similar variation in size. Depending upon size, zip codes and counties were used individually or combined to create 61 geographic areas. Because the local health district is the primary seat of community public health decision- making in Utah, small areas were geographically constrained so that their boundaries would not cross local health district boundaries. Most multi-county Utah health districts contained more than one small area.

Adjacency: There will be numerous zip codes within the political boundaries decided by you, and the tie can be broken by selecting contiguous zip codes for grouping. The assumption is that the adjacent areas are likely to be similar with respect to various community level attributes. But again there might be multiple zip codes qualifying the criterion of adjacency, whereas you may need to combine fewer, to stay within the population size criterion for small areas. The next criterion would have to be decided at this stage.

In combining zip codes to define **Utah small areas**, in all but two cases, only contiguous zip codes were combined. With only one exception, sub-county small areas were wholly contained within individual counties and were not combined with zip code areas in neighboring counties. Whenever possible, the areas were designed to conform to established political boundaries of cities and towns.

Family, household or per capita Income: Income is an important variable in determining the quality of and access to healthcare. Studies have shown that socioeconomic factors are significant determinants of the variation in both medical and surgical discharge rates (McLaughlin et al. 1989). To break the tie between the contiguous zip codes, you may use the average (median) household income, family income, or per capita income. Of these, per capita income is more refined measure. Family credit and income Support data have been recommended by Carr-Hill et al. (20002) in building small area analysis models. After considering the income, if there are still ties in competing zip codes for a certain small area, you can continue using the other characteristics. For instance, consider the natural landmarks such as mountains, or man made structures such as freeways etc. in making your decisions to combine zip codes. Other possible zip code characteristics that can be used would include median age, some measures of education level, etc.

In case of a tie between zip codes in deciding a small area in **Utah**, median per capita annual income levels of each zip code area were used to ascertain homogeneity with respect to socioeconomic status within a small area. After addressing the criteria listed above, there still remained areas whose boundaries had not been set.

Perceptions of the team and consensus building among local representatives: In addition to the criteria discussed earlier, discussing the small area scheme with the workgroup, professionals, and other experts on the characteristics of the various areas in state, and making appropriate adjustments is highly advised.

Building or obtaining consensus for a methodology was a unique feature of **Utah's methodology**. In addition to the technical solution, the "buy-in" across programs and stakeholders was obtained so that there could be comparability in measures across programs. Some large urban counties were subdivided into many small areas. Consequently there were ties among zip codes even when income levels were considered. In those cases, zip codes were combined based on the perception of the team involved in the small area analysis about similarity of the populations. In some cases, county or district health departments were consulted. The resulting draft small area design was then submitted to local representatives, primarily in areas where subjective criteria had been used to combine zip code areas. The local representatives (10 of the 12 Utah local health officers, and 26 city officials selected from the directory of the Utah League of Cities and Towns) were provided a map of their locality showing the proposed small area boundaries and asked to consider whether the combined zip code areas were similar in terms of lifestyle and demographic characteristics. Several changes were made based on their recommendations.

C. Applications

The importance of small area analysis is due primarily to the variety of its applications, some of which are listed below:

Small area analysis is essential for states for their community level planning and assessment.

Small area analysis is commonly used to understand disparities with respect to mortality and morbidity across communities/small areas to use as a base for planning the delivery of health services (Ubido and Ashton, 1993).

Small area analysis has been used for understanding health care services needs within specific geographic areas. These analysis are often performed for exploring the effects of supply factors versus population needs regarding the distribution of health services (Paul-Shaheen, Clark, & Williams, 1987; Wennberg, 1987; Wilson & Tedeschi, 1984).

Studies also use small area analysis to examine the effect of supply factors on health care patterns (Connell, Day and LoGerfo 1981), and variation in use of services (Gittelsohn and Powe, 1995)

Understanding health situation sometimes requires linking information from various data sources. When matching individuals across datasets is not possible because of absence of a unique identifier, or identifiers allowing probabilistic linkage (Shah, 2000), small area analysis framework enables linkage of health data from multiple sources at a community level.

Studies use small area analysis for investigating patterns of diseases that are typically affected by community level environmental risks (e.g., Lang & Polansky, 1994), including injury surveillance (Durkin et al., 1994; Borell et al 2002), abortion rates (Ubido J, Ashton 1993), homicide victimization (Gjelsvic et al. (2004), Asthma (Moudgil, Marshall, and Honeybourne, 2000; Ray et al. 1998), and malaria (Kleinschmidt 2002).

Studies have also used small area analysis for detecting clustering of disease occurrence (Hole and Lamont, 1992). Small area analysis has also been used to demonstrate that the socio-economic status of residential areas was associated with mortality (Waitzman & Smith, 1998; Gould, Davey and LeRoy, 1989) incidence rates for colorectal cancer (Haining, Wise Blake, 1994) and with primary cesarean section rates (Gould, Davey & Stafford, 1989).

Other applications of small area analysis in public health have included study of ethnic enclaves (Moudgil, Marshall, and Honeybourne, 2000) and population groups with special needs (Andrews, Kemer, Zauber et al., 1994; Kleinman, 1977).

In sum, this document has been written as a general guideline for the states to develop a small area analysis scheme to be applied to the healthcare data. It draws on the author's experience in developing small area analysis at the Utah Department of health. Small areas are areas that are large enough to have a sufficient number of events of interest to yield stable rates, yet they are small enough to unmask variations in the rates and still convey a sense of community. Creation of small area scheme is necessary because the existing geographic units typically used in the public health data analysis (e.g. county or local health district) are inadequate for community level surveillance. Developing a small area analysis scheme involves several steps that are listed in this document. It merits a mention that different states may face different set of challenges in developing small area analysis. The applications of small area analysis are also briefly listed in this paper.

D. References

Andrews H F, Kemer J F, Zauber A G, Mandelblatt J, Pittman J, and Struening E. (1994). Using census and mortality data to target small areas for breast, colorectal, and cervical cancer screening. American *Journal of Public Health*. 84: 56-61.

American Public Health Association (1991) Healthy Communities 2000: Model Standards. Guidelines for community attainment of the year

2000 National Health Objectives (3rd Ed.). Washington DC: American Public Health Association.

APEXPH Steering Committee (1991) APEXPH: Assessment Protocol for Excellence in Public Health. Washington DC: National Association of County Health Officials.

Borrell C, Rodríguez M, Ferrando J, Brugal M T, Pasarín M I, Martínez V, and Plasència A. (2002). Role of individual and contextual effects in injury mortality: new evidence from small area analysis. *Injury Preview*. 8: 297 - 302.

Carr-Hill RA, Jamison JQ, O'Reilly D, Stevenson MR, Reid J, and Merriman B. (2002). Risk adjustment for hospital use using social security data: cross sectional small area analysis. *British Medical Journal*. 324: 390 - 92

Connell FA, Day RW, and LoGerfo JP. (1981). Hospitalization of Medicaid children: analysis of small area variations in admission rates *American Journal of Public Health*. 71: 606 - 613.

Durkin MS, Davidson LL, Kuhn L, O 'Connor P, and Barlow B (1994).
Low-income neighborhoods and the risk of severe pediatric injury: a small-area analysis in northern Manhattan *American Journal of Public Health*. 84: 587 - 592.

Gittelsohn A, and Powe NR. 1995. Small area variations in health care delivery in Maryland. *Health Services Research*. 30(2):295-317

Gjelsvik A, Zierler S, and Blume J. (2004). Homicide risk across race and class: a small-area analysis in Massachusetts and Rhode Island *Journal of Urban Health*. 81: 702 - 718.

Goodman D C and Green G R. (1996). Assessment tools: Small area analysis. *American Journal of Medical Quality.* 11(1), S 12-14.

Gottlieb DJ, Beiser AS, and O'Connor GT. (1995). Poverty, race, and medication use are correlates of asthma hospitalization rates: A small area analysis in Boston. *Chest*. 108: 28 - 35.

Gould JB, Davey JB, and Stafford RS. (1989). Socioeconomic Differences in Rates of Cesarean Section. *New England Journal of Medicine*. 321(4), 233-239.

Haggard LM, Shah GH, and Rolfs RT. 1999. Assessing health status: establishing geographic areas for small area analysis in Utah. *Utah's Health: An Annual Review Volume.* V 1997-98 Pp 18-35

Haining R., Wise S, and Blake M. (1994). Constructing regions for small area analysis: material deprivation and colorectal cancer. *Journal of Public Health Medicine*. 6: 429 - 438.

Hole DJ, and Lamont DW. (1992). Problems in the interpretation of small area analysis of epidemiological data: the case of cancer incidence in the West of Scotland *Journal of Epidemiology and Community Health*. 46: 305 - 310.

Joines, J D., Hertz-Picciotto I, Carey TS, Gesler W, and Suchindran C. (2003). A spatial analysis of county-level variation in hospitalization rates for low back problems in North Carolina. *Social Science & Medicine*. 56(12):2541-53.

Kahn HA., and Sempos CT. (1989). *Statistical Methods in Methodology*. New York:Oxford University Press

Kazandjian VA, Hudson J. (1990). Standards and norms in small area analysis. *Maryland Medical Journal*. (3):249-51.

Kazandjian VA, Durance PW, Schork MA. (1989). The extremal quotient in small-area variation analysis. *Health Services Research*. 24(5):665-84.

Kleinman JC. (1977). Age-adjusted mortality indexes for small areas: Applications to health planning. *American Journal of Public Health*. 67:834-840.

Kleinschmidt B, Sharp I, Mueller and Vounatsou P. (2002). Rise in malaria incidence rates in South Africa: A Small-area spatial analysis of variation in time trends *American Journal of Epidemiology*.155 (3): 257 - 264.

Lang D M and Polansky M. (1994). Patterns of asthma mortality in Philadelphia from 1969 to 1991. *The New England Journal of Medicine*. 331: 1542-1546.

McLaughlin CG, Normolle DP, Wolfe RA, McMahon LF Jr, and Griffith JR. (1989). Small-area variation in hospital discharge rates. Do socioeconomic variables matter? Medical Care. 27(5):507-21.

Moudgil H, Marshall T, and Honeybourne D. (2000). Asthma education and quality of life in the community: a randomized controlled study to evaluate the impact on white European and Indian subcontinent ethnic groups from socioeconomically deprived areas in Birmingham, UK. *Thorax*. 55(3): 177 - 183.

Parchman ML. (1995). Small area variation analysis: a tool for primary care research. *Family Medicine*. 27(4):272-6.

Paul-Shaheen P, clark J D, & Williams, D. (1987). Small area analysis: a review and analysis of the North American literature. *Journal of Health Politics, Policy and Law*. 12: 741-809.

Ray NF, Thamer M, Fadillioglu B, and Gergen PJ. (1998). Race, income, urbanicity, and asthma hospitalization in California: a small area analysis. Chest. 113: 1277 - 1284.

Shwartz M, Ash AS, Anderson J, Iezzoni LI, Payne SM, and Restuccia JD. (1994). Small area variations in hospitalization rates: how much you see depends on how you look. *Medical Care*. 32(3):189-201.

Shah GH. (2000). Probabilistic data linkage in public health research *Utah's Health: An Annual Review*. 5:128-129

Stano M. (1993). Evaluating the policy role of the small area variations and physician practice style hypotheses**.** *Health Policy*. 24(1):9-17.

Spasoff R A, Strike C J, Nair R C, Dunkley G C, and Boulet J R. (1996). Small group estimation for public health. *Canadian Journal of Public Health*, 87(2), 130-134.

Ubido J, and Ashton J. (1993). Small area analysis: abortion statistics *Journal of Public Health Medicine*. 15: 137 - 143.

Waitzman N J, and Smith K R. (1998). Phantom of the area: poverty area residence and mortality in the United States. *American Journal of Public Health*. 88(5), 1-4.

Wennberg JE. (1987). Population illness rates do not explain population hospitalization rates. *Medical Care*. 25, 354-359.

Wilson, P., & Tedeschi, P. (1984). Community Correlates of Hospital Use. *Health Services Research*. 19, 333-355.

Administrative Data and Disease Surveillance: An Integration Toolkit

NAHDO-CDC Cooperative Agreement Project
CDC Assessment Initiative

By: Barbara Rudolph, PhD
Robert Davis, MS

The National Association of
Health Data Organizations

ADMINISTRATIVE DATA AND DISEASE SURVEILLANCE: AN INTEGRATION TOOLKIT

A. 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, and 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.

B. Administrative Data/Surveillance Data: Similarities and Differences?

Currently, 48 states collect 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 more recently, the 837 professional claim formats. Data editing programs are used to improve the quality of the data that are 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) 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 (1999) 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.

C. What is Integration?

In a formal sense, data integration is the process of combining two or more data sets together, or movement of data between two co-existing systems, for sharing and analysis, in order to support information management. (Informatica, 2005; GTU, 2005). For the purposes of this paper, we are defining integration as data sharing between administrative and clinical data systems. 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.

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 workforce 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.

D. 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 wellbeing 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(e.g. threats to use Bacillas anthvacis bacterium to spread anthrax) 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 and 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 are 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[11] the other.

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

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 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.

E. 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.

F. 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.

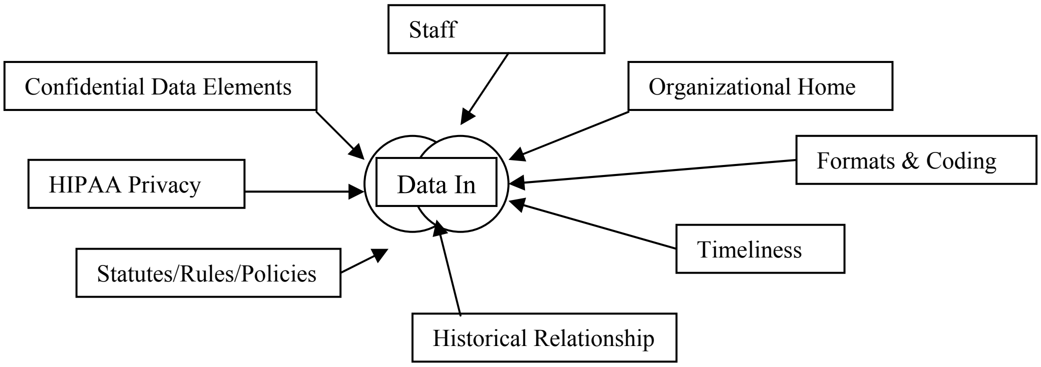

Figure 1. Barriers to Integration

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[12]. It will require a thorough analysis of the necessary data elements and the potential restrictions on use whether state or federally required.

[12] 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

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.

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[13] 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 impossible given substantial differences in penalties for releasing information, and policies on re-release of information. The database history,

[13] 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.

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.

G. 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 are needed and how are they 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.

H. 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 are 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 are 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.

I. 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.

J. 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.

K. 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.

K. 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

L. 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.

References

GTU. (2005). Data Warehouse Glossary. University Information Services, Georgetown University online at: http://uis.georgetown.edu/departments/eets/dw/GLOSSARY0816.html Retrieved Nov. 10, 2005.

Hebert, P. L., L. S. Geiss, E. F. Tierney, M. M. Engelgau, B. P. Yawn, and A. M. McBean. (1999). "Identifying Persons with Diabetes Using Medicare Claims Data."*American Journal of Medical Quality* 14 (6): 270–7.

Informatica, (2005). Technical Glossary. Online at http://www.informatica.com/solutions/resource_center/glossary/default.htm Retrieved Nov. 10, 2005.