Statistical Issues in Interactive Web-based Public Health Data Dissemination Systems

MICHAEL A. STOTO

WR-106
October 2003
Prepared for the National Association of Public Health Statistics and Information Systems
EXECUTIVE SUMMARY

State- and community-level public health data are increasingly being made available on the World Wide Web for the use of professionals and the public. The goal of this paper is to identify and address the statistical issues associated with these interactive data dissemination systems. The analysis is based on telephone interviews with 14 individuals in five states involved with the development and use of seven distinct interactive web-based public health data dissemination systems, as well as experimentation with the systems themselves.

Interactive web-based systems offer state health data centers an important opportunity to disseminate data to public health professionals, local government officials, and community leaders, and in the process raise the profile of health issues and involve more people in community-level decision making. The primary statistical concerns with web-based dissemination systems relate to the small number of individuals in the cells of tables when the analysis is focused on small geographic areas or in other ways. In particular, data for small population groups can be lacking in statistical reliability, and also can have the potential for releasing confidential information about individuals. These concerns are present in all statistical publications, but are more acute in web-based systems because of their focus on presenting data for small geographical areas.
Small numbers contributing to a lack of statistical reliability

One statistical concern with web-based dissemination systems is the potential loss of statistical reliability due to small numbers. This is a concern in all statistical publications, but it is more acute in web-based systems because of their focus on presenting data for small geographical areas and other small groups of individuals.

There are a number of statistical techniques that interactive data dissemination systems can use to deal with the lack of reliability resulting from small cell sizes. Aggregation approaches can help, but information is lost. Small cells can be suppressed, but even more information is lost. (The best rationale for numerator-based data suppression is confidentiality protection, not statistical reliability.) In general, approaches that use statistical approaches to quantify the uncertainty (such as confidence intervals and the use of \( c^2 \) tests), or to smoothing, or small area model-based estimation, should be preferred to options that suppress data or give counts but not rates.

Small numbers and confidentiality concerns

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. The definition of “small” varies by state, and often by dataset. This approach often results in a substantial loss of information and utility.

Statisticians in a number of state health data centers have recently reconsidered data suppression guidelines currently in use and have developed
creative and thoughtful new approaches, as indicated above. Their analyses, however, have not been guided by theory or statistical and ethical principles, and have not taken account of extensive research on these issues and development of new methods that has taken place in the last two decades. Government and academic statisticians, largely outside of public health, have developed a variety of “perturbation” methods such as “data swapping” and “controlled rounding” that can limit disclosure risk while maximizing information available to the user. The Census Bureau has developed a “confidentiality edit” to prevent the disclosure of personal data in tabular presentations. The disclosure problem can be formulated as a statistical decision problem that explicitly balances the loss that is associated with the possibility of disclosure and the loss associated with non-publication of data. Such theory-based and principled approaches should be encouraged.

**Concept validity and data standards**

Statisticians have been concerned ever since computers were introduced that the availability of data and statistical software would lead untrained users to make mistakes. While this is probably true to some extent, restricting access to data and software is not likely to succeed in public health. The introduction of interactive web-based dissemination systems, on the other hand, should be seen as an important opportunity to develop and extend data standards in public health data systems.

Web-based dissemination systems, because they require that multiple data systems be put into a common format, present opportunities to disseminate
appropriate data standards and to unify state data systems. Educational efforts building on the dissemination software itself, as well as in more traditional settings, are likely to be more effective in reducing improper use of data than restricting access. For many users, such training will need to include content on using public health data, not just on using web-based systems. The development of standard reports for web-based systems can be an effective means for disseminating data standards.

Data validation

No statistical techniques can guarantee that there will be no errors in web-based data systems. Careful and constant checking of both the data and the dissemination system, as well as a policy of releasing the same data files to all users, however, can substantially reduce the likelihood of errors. Methods for validation should be documented and shared among states.

The development of web-based dissemination systems is an opportunity to implement data standards rather than a problem to be solved. Efforts to check the validity of the data for web dissemination purposes may actually improve overall data quality in state public health data systems.

General comments

The further development and use of web-based data dissemination systems will depend on a good understanding of the systems’ users and their needs. System designers will have to balance between enabling users and protecting users from themselves. Systems will also have to develop ways to
train users not only in how to use the systems themselves, but also on statistical issues in general and the use of public health data.

Research to develop and implement new statistical methods, and to better understand and address users’ needs, is a major investment. Most states do not have the resources to do this on their own. Federal agencies, in particular through CDC’s Assessment Initiative, could help by enabling states to share information with one another, and by supporting research on the use of new statistical methods and on data system users.
INTRODUCTION

State- and community-level public health data are increasingly being made available on the World Wide Web for the use of professionals and the public. Although most data of this sort currently available are simply static presentations of reports that have previously been available in printed form, interactive web-based systems are increasingly common (Friedman et al, 2001).

The goal of this paper is to identify and address the statistical issues associated with interactive web-based state health data dissemination systems. This will include assessing the current data standards, guidelines, and/or best practices used by states in their dissemination of data via the Web for both static presentation of data and interactive querying of data sets and analyzing the statistical standards and data dissemination policies, including practices to ensure compliance with privacy and confidentiality laws. Many of the same statistical issues apply to public health data however published, but interactive web-based systems make certain issues more acute. In addition, identifying and addressing these issues for interactive systems may also lead to overall state health data system improvement.

This analysis is based on telephone interviews with 14 individuals in five states involved with the development and use of seven distinct interactive web-based public health data dissemination systems, as well as experimentation with the systems themselves. All but one of the systems are currently in operation, but most are constantly being updated. The interviewees and information on the sites appears in Appendix A. The choice of these individuals and states was not intended to be exhaustive or representative, but to bring out as many statistical
issues as possible. In addition, a preliminary draft of this paper was circulated for comment and was discussed at a two-day workshop at Harvard School of Public Health in August, 2002; attendees are listed in Appendix B. The current draft reflects comments by e-mail and at the workshop, but the analysis and conclusions are the author’s, as well as any errors that may remain.

This paper begins with a background section that addresses the purposes, users and benefits of interactive data dissemination systems, systems currently in place or being developed, and database design as it affects statistical issues. The body of the paper is organized around four substantive areas: (1) small numbers contributing to a lack of statistical reliability; (2) small numbers leading to confidentiality concerns; (3) concept validity and data standards, and (4) data validation. The paper concludes with a summary and conclusions. A glossary of key terms appears in Appendix C.

BACKGROUND

Purposes, users, and benefits of interactive data systems

Interactive web-based data dissemination systems in public health have been developed for a number of public health assessment uses. One common use is to facilitate the preparation of community-level health profiles. Such reports are consistent with Healthy People 2010 (DHHS, 2000), and are increasingly common at the local/county level. In some states, they are required. This movement reflects the changing mission of public health from direct delivery of personal health care services to assessment and policy development (IOM,
The reports are used for planning and priority setting as well as for evaluation of community-based initiatives.

Minnesota, for instance, will use its interactive dissemination system to reshape the way that state and county health departments do basic reports by facilitating, and hence encouraging, the use of certain types of data. The system is intended to provide better and more current information to the public than is available in the current static system, in which data are updated only every two years.

From another perspective, the purpose of web-based dissemination systems is to enable local health officials, policy makers, concerned citizens, and community leaders who are not trained in statistics or epidemiology to participate in public health decision-making. Because many of these users are not experienced data users, some systems are designed to help users find appropriate data. MassCHIP, for instance, was designed with multiple ways into datasets so users are more likely to “stumble upon” what they need. Users can search, for instance, using English-language health problems lists and Healthy People objectives, as well as lists of datasets.

Web-based dissemination systems are also a natural outgrowth of the activities of state health data centers. The systems allow users to prepare customized reports (their choice of comparison areas, groups of ICD codes, age groups, and so on). So in addition to making data available to decision makers and the public, they also facilitate work already done by state and local public health officials and analysts. This includes fulfilling data requests to the state data center as well as supporting statistical analyses done by subject area
experts. States have seen substantial reduction in the demand on health statistics staff for data requests. In at least one case, the system itself has helped to raise the profile of the health department with legislators.

Interactive web data systems are also being used to detect and investigate disease clusters and outbreaks. This includes cancer, infectious diseases, and, increasingly, bioterrorism. Interactive web systems are also being used, on a limited basis, for academic research, or at least for hypothesis generation. The software that runs some of these systems (as opposed to the state health data that are made available through it) has also proven useful for research purposes. Nancy Krieger at the Harvard School of Public Health, for instance, is using VistaPH to analyze socio-economic status data in Massachusetts and Rhode Island, and others are using it in Duval County, Florida and Multnomah County, Oregon.

Some states are also building web-based systems to bring together data from a number of health and social service programs and make them available to providers in order to simplify and coordinate care and eligibility determination. Such systems can provide extremely useful statistical data, and in this sense are included in this analysis. The use of these systems for managing individual patients, however, is not within the scope of this paper.

Reflecting the wider range of purposes, the users of web-based data systems are very diverse. They include local health officials, members of boards of health, community coalitions, as well as concerned members of the public. Employees of state health data centers, other health department staff, and employees of other state agencies; hospital planners and other health service
administrators, public health researchers and students of public health and other health fields also use the data systems. 

These users range from frequent to occasional. Frequent users can benefit from training programs and can use more sophisticated special purpose software. Because most users only use the system occasionally, there is a need for built-in help functions and the like. Tennessee’s system, for instance, is colorful and easy to use. Elementary school students up to graduate students in community health courses have used it for class exercises.

Because of the breadth of uses and users, the development of a web-based dissemination system can lead to consideration and improvement of data standards and to more unification across department data systems. This happens by encouraging different state data systems to use common population denominators, consistent methods, such as for handling missing data, consistent data definitions, for example for race/ethnicity and common methods for age adjustment (to the same standard population) and other methods, such as confidence intervals.

**Current web-based public health data dissemination systems**

In support of the wide variety of uses and users identified above, current public health web-based data dissemination systems include many different kinds of data. Each of the following kinds of data is included in at least one of the seven data systems examined for this study. Reflecting the history of state health data centers, vital statistics are commonly included. Most systems also include census or other denominator data needed to calculate population-based
rates. In support of community health assessment initiatives, web-based dissemination systems also typically include information related to *Healthy People 2010* (DHHS, 2000) measures or their state equivalents, and links to HRSA’s Community Health Status Indicators (HRSA, undated).

Systems also commonly include data “owned” by components of the public health agency outside the state data center, and sometimes by other state agencies. Web-based dissemination systems, for instance, typically include epidemiologic surveillance data on infectious diseases, including HIV/AIDS, and, increasingly, bioterrorism. Cancer registry data are included in some systems. Some systems include health services data based on hospitalization, such as average length of stay and costs, as well as Medicaid utilization data. One system includes data from outside the health department on TANF and WIC services.

Although much of the data covered by web-based dissemination systems is based on individual records gathered for public health purposes, such as death certificates and notifiable disease reports, population-based survey data are also included. Data from a state’s Behavioral Risk Factor Surveillance System (BRFSS) (CDC, undated), youth behavioral risk factor and tobacco surveys where available, and others, are commonly included.

Demographic detail in web dissemination systems generally reflects what is typically available in public health data sets and what is used in tabulated analyses: age, race, sex, and sometimes indicators of socioeconomic status. Definitions of these variables and how they are categorized frequently vary across the data sets available in a single state system.
The geographic detail in web-based dissemination systems, however, is substantially greater than is typically available in printed reports. State systems typically have data available for each county or, in New England states, town. Some of the state systems also have data available for smaller areas in large cities. Missouri’s MICA makes some health services data available by Zip code. Some of the systems allow users to be flexible in terms of disaggregation. The basic unit in MassCHIP is the city/town, but the system allows users to group these units into community health areas, HHS service areas, or user-defined groups. The VistaPHs and EpiQMS systems in Washington allow user-defined groups based on census block.

This geographical focus, first of all, is designed to make data available at the level of decision-making, and to facilitate the participation of local policy makers and others in health policy decisions. This focus also allows public health officials to develop geographically and culturally targeted interventions. In Washington, for instance, a recent concern about teen pregnancy led public health officials to identify the counties, and then the neighborhoods, with the highest teen fertility rates. This analysis led them to four neighborhoods, two of which were Asian where teen pregnancy is not considered a problem. They were then able to focus their efforts in the two remaining neighborhoods. In the future they anticipate using the system to support other surveillance activities as well as outbreak investigations.

Although the combination of demographic and geographic variables in theory allows for a great degree of specificity, in actual practice the combination
of these variables is limited by database design, statistical reliability, and confidentiality concerns, as discussed below.

Data availability and state priorities drive what is included in web-based dissemination systems. According to John Oswald, for instance, Minnesota’s health department has three priorities – bioterrorism, tobacco, and disparities – so the system is being set up to focus on these. Data availability is a practical issue; it includes whether data exist at all, are in a suitable electronic form, come with arrangements that allow or prohibit dissemination, and whether the data are owned by the state health data center.

State public health data systems are also an arena in which current statistical policy issues are played out, and this has implications for database content and design. Common concerns are the impact of the new Health Insurance Portability & Accountability Act of 1996 (HIPAA) regulations regarding the confidentiality of individual health information (Gostin, 2001), the recent change in federal race/ethnicity definitions, the adoption of the Year 2000 standard population by the National Center for Health Statistics (Anderson and Rosenberg, 1998), and surveillance for bioterrorism and emerging infectious diseases.

In the future, web-based dissemination systems will likely be expanded to include more data sets. Some states are considering using these systems to make individual-level data available on a restricted basis for research purposes. States are also considering using these systems to make non-statistical information (e.g. breast cancer fact sheets, practice guidelines, information on
local screening centers) available to community members, perhaps linked to data requests on these subjects.

**Database design**

Current web-based dissemination systems range from purpose-built database software to web-based interfaces to standard, high-powered statistical software such as SAS or GIS systems such as ESRI Map Objects that resides on state computers. System development has been dependent on the statistical, information technology, and Internet skills available in (and to) state health data centers. Missouri and Massachusetts built their own systems. Washington adopted a system built by a major local health department, Seattle-King County. Tennessee contracted with a university research group with expertise in survey research and data management. Not surprisingly, the systems have evolved substantially since they were first introduced in 1997 due to changes in information technology, and Internet technology, and the availability of data in electronic form.

The designers of web-based dissemination systems in public health face two key choices in database development. As discussed in detail below, these choices have statistical implications in terms of validity checking, choice of denominator, data presentation (e.g. counts vs. rates vs. proportions, etc.), ability to use sophisticated statistical methods, and user flexibility.

First, systems may be designed to prepare analyses from individual-level data “on the fly” – as requested – as in MassCHIP, or to work with preaggregated data (Missouri’s MICA) or pre-calculated analytical results (Washington’s
EpiQMS). “On the fly” systems obviously have more flexibility, but the time it takes to do the analyses may discourage users. This time can be reduced by pre-aggregation. Rather than maintaining a database of individual-level records, counts of individuals who share all characteristics are kept in the system. Different degrees of preaggregation are possible. At one extreme, there are static systems in which all possible tables and analyses are prepared in advance. At the other, all calculations are done using individual-level data. In between, a system can maintain a database with counts of individuals who share characteristics. The more characteristics that are included, the more this approaches an individual-level system.

At issue here is the degree of user control and interaction. Static systems can deliver data faster, but are less flexible in what can be requested and may limit the user in following up leads that appear in preliminary analyses. The preprocessing step, however, can provide an opportunity for human analysts to inspect tables and ensure that statistical analyses supporting more complex analyses are reasonable. EpiQMS, for instance, uses an “accumulated record” database – all of the calculations have been done in advance – which allows for greater speed and a user-friendly design. It also allows the system administrators to look at the data and see if it makes sense, and also to identify and fix problems based on human intelligence.

The second major design choice is between a server-resident data and analytic engine vs. a client-server approach. In a server-resident system, the web serves as an interface that allows users to access data and analytic tools that reside on a state health data center server. The only software that the user
needs is a web browser. In a client-server approach, special purpose software on the user's computer accesses data on the state's computer to perform analyses. Client-server software allows for greater flexibility (for example, users can define and save their own definitions of geographical areas), but the necessity of obtaining the client software in advance can dissuade infrequent users.

Systems can, and do, combine these approaches. MassCHIP, for instance, uses client-server software to do analyses on the fly, but makes a series of predefined “Instant Topics” reports available through the web. Tennessee’s HIT system uses a combination of case-level and prepared analyses. Its developers would like more case-level data because it is more flexible, but these analyses are hard to program, and resources are limited. In the end, users care more about data than datasets, so an integrated front end that helps people find what they need is important.

Although all of the systems include some degree of geographical data, they vary in the way that these data are presented. Washington’s EpiQMS and Tennessee’s HIT systems feature the use of data maps, which are produced by commercial GIS software. According to Richard Hoskins, spatial statistics technology has finally arrived, and the EpiQMS system makes full use of it. The systems also differ in the availability of graphical modes of presentation such as bar and pie charts.

The design of web-based dissemination systems should, and does, represent the diversity of the users as discussed above. A number of systems, for instance, have different levels for types of users. Users differ with respect to
their statistical skills, their familiarity with computer technology, and their substantive expertise, and there is a clear need (as discussed in more detail below) for education, training, and on-line help screens that reflect the different skills and expertise that the users bring to the systems.

**ANALYSIS AND RECOMMENDATIONS**

Small numbers contributing to a lack of statistical reliability

Because of their focus on community-level data, web dissemination systems for public health data eventually, and often quickly, get to the point where the numbers of cases available for analysis become too small for meaningful statistical analysis or presentation. It is important to distinguish two ways in which this can happen.

First, in statistical summaries of data based on case reports oftentimes the expected number of cases is small, meaning the variability is relatively high. In statistical summaries the number of reported cases \( x \) typically forms the numerator of a rate \( p \), which could be the prevalence of condition A per 1,000 or 100,000 population, the incidence of condition B per 100,000 residents per year, or other similar results. Let the base population for such calculations be \( n \). There is variability in such rates from year to year and place to place because of the *stochastic variability* of the disease process itself. That is, even though two communities may have the same, unchanging conditions that affect mortality, and 5 cases would be expected in each community every year, the actual number in any given year could be 3 and 8, 6 and 7, and so on, simply due to chance.
The proper formula for the variance of such rates depends on the statistical assumptions that are appropriate, typically binomial or Poisson. When $p$ is small, however, the following formulas hold approximately:

1. \[ \text{Var} (x) = np \]
2. \[ \text{Var} (p) = \text{Var} (x/n) = p/n \]

Analogous formulae are available for more complex analyses, such as standardized rates, but the fundamental relationship to $p$ and $n$ is similar.

Since the expected value of the number of cases, $x$, is also equal to $np$, the first formula implies that the standard deviation of $x$ equals the square root of its expected value. If the expected number of cases is, say, 4, the standard deviation is $\sqrt{4} = 2$. If the rate were to go up by 50% so that the expected number of cases became 6, that would be only about 1 standard deviation above the previous mean, and such a change would be difficult to detect. In this sense, when the number (or more precisely the expected number) of cases is small, the variability is relatively high.

The second formula, on the other hand, reminds us that the population denominator is also important. In terms of rates, a rate calculation based on 4 events is far more precise if the population from which it is drawn is 10,000 than if it is 100. In the first case $p = 4/10,000 = 0.0004$ and the standard deviation is $\sqrt{0.0004/10,000} = 0.02/100 = 0.0002$. In the second case $p = 4/100 = 0.04$ and the standard deviation is it is $\sqrt{0.04/100} = 0.2/10 = 0.02$. In addition, in two situations leading to the same calculated rate of $p$, the one with the larger $n$ is also more precise. For instance, $400/10,000$ and $4/100$ both yield $p = 0.04$, but
the standard deviation of the first is $\div 0.04/10,000 = 0.2/100 = 0.002$ and the second is $\div 0.04/100 = 0.2/10 = 0.02$. Table 1 illustrates these points.

Table 1. Small numbers and statistical reliability examples

<table>
<thead>
<tr>
<th>$x$ = 4, $n$ = 100 or 10,000</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$ = 4</td>
<td>$n$ = 100</td>
<td>SD($p$) = $\div 0.04/100 = 0.02$</td>
<td>SD/$p$ = 0.02/0.04 = 0.5</td>
</tr>
<tr>
<td>$x$ = 4</td>
<td>$n$ = 10,000</td>
<td>SD($p$) = $\div 0.0004/10,000 = 0.0002$</td>
<td>SD/$p$ = 0.0002/0.0004 = 0.5</td>
</tr>
<tr>
<td>$p$ = 0.04, $n$ = 100 or 10,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p$ = 0.04</td>
<td>$n$ = 100</td>
<td>SD($p$) = $\div 0.04/100 = 0.02$</td>
<td>SD/$p$ = 0.02/0.04 = 0.5</td>
</tr>
<tr>
<td>$p$ = 0.04</td>
<td>$n$ = 10,000</td>
<td>SD($p$) = $\div 0.04/10,000 = 0.002$</td>
<td>SD/$p$ = 0.002/0.04 = 0.05</td>
</tr>
</tbody>
</table>

Another way of looking at these examples is in terms of relative reliability, which is represented by the standard deviation divided by the rate ($SD/p$). As the top of Table 1 illustrates, the relative standard deviation depends only on the numerator; when $x$ is 4, $SD/p$ is 0.5 whether $n$ is 100 or 10,000. When $p$ is held constant, however, as in the bottom lines of Table 1, the relative standard deviation is smaller when $n$ is larger.

The second situation that yields small numbers in community public health data is the calculation of rates or averages based on members of a national or state sample that reside in a particular geographical area. A state’s BRFSS sample, for instance, may include 1,000 individuals chosen through a scientific sampling process. The size of the sample in county A, however, may be 100, or 10, or 1. This presents two problems.

First, in a simple random sample, *sampling variability* in a sample of size $n$ can be described as

(3) \[ Var(p) = p(1-p)/n \]
for proportions. Note that if $p$ is small the variance is approximately $p/n$ as above, but the $n$ is the size of the sample, not the population generating cases. When $p$ is small the relative standard deviation is approximately equal to $\sqrt{(p/n)/n} = \sqrt{(pn)}$. Since $pn$ equals the numerator, $x$, two samples with the same number of cases will have the same relative standard deviation, as above.

For sample means,

$$\text{(4) } \text{Var}(\bar{x}) = \frac{s}{n}$$

where $s$ is the standard deviation of the variable in question. In both (3) and (4), the sampling variability is driven by the size of the sample, which can be quite small. The variance of estimates based on more complex sampling designs is more complex, but depends in the same fundamental way on the sample size, $n$.

The second problem is that while sampling theory ensures an adequate representation of different segments of the population in the entire sample, a small subsample could be unbalanced in some other way. If a complex sampling method is used for reasons of statistical efficiency, some small areas may, by design, have no sample elements at all.

The major conclusion of this analysis, therefore, is that the variability of rates and proportions in public data is generally inversely proportional to $n$, which can be quite small for a given community. For epidemiologic or demographic rates the variability is stochastic, and $n$ is the size of the resident population generating the cases. For proportions or averages based on a random sample, $n$ is the size of the sample in the relevant community. For a proportion, the relative standard deviation is proportional to the expected count.
When the primary source of variability is stochastic variation due to a rare health event, the “sample size” cannot be changed. When the data are generated by a sample survey, it is theoretically possible to increase the sample size to increase statistical precision. Sampling theory tells us, however, the absolute size of the sample, \( n \), rather than the proportion of the population of the target population that is sampled, drives precision. If a sample of 1,000 is appropriate for a health survey in a state, a sample of 1,000 is needed in every community in the state (whether there are a million residents or 50 thousand) to get the same level of precision at the local level.

Although sample size for past years is fixed, states can and have increased the sample size of surveys to allow for more precise community-level estimates. Washington, for instance, “sponsors” three counties each year to increase the BRFSS sample size to address this problem. In the late 1990s, Virginia allocated its BRFSS sample so that each of the health districts had about 100 individuals, rather than proportional to population size. Combining three years of data, this allowed for compilation of county-level health indicators for the Washington metropolitan area (Metropolitan Washington Public Health Assessment Center, 2001). California has recently fielded its own version of the National Health Interview Survey with sufficient sample size to make community-level estimates throughout the state (UCLA, 2002). Increasing sample size, however, is an expensive proposition.

The typical way that states resolve the problem of small numbers is to increase the effective \( n \) by aggregating results over geographic areas or multiple years. The drawbacks to this approach are obvious. Aggregating data from
adjacent communities may hide differences in those communities, and make the data seem less relevant to the residents of each community. Combining data for racial/ethnic groups, rather than presenting separate rates for Blacks and Hispanics, can mask important health disparities. Aggregating data over time requires multiple years of data, masks any changes that have taken place, and the results are “out of date” since they apply on average to a period that began years before the calculation is done. Although many public health rates may not change all that quickly, simply having data that appears to be out of date affects the credibility of the public health system. Depending on the size of the community, modest amounts of aggregating may not be sufficient to increase $n$ to an acceptable level, so data are either not available or suffer even more from the problems of aggregation.

Another typical solution is to suppress results (counts, rates, averages) based on fewer than $x$ observations. This approach is sometimes referred to as the “rule of 5” or the “rule of 3” depending on the value of $x$. In standard tabulations, such results are simply not printed. In an interactive web-based data dissemination system, the software would not allow such results to be presented to the user. The user only knows that there were fewer than $x$ observations, and sometimes more than 0. Rules of this sort are often motivated on confidentiality grounds (see the following section), so $x$ can vary across and within data systems, and typically depend on the subject of the data rather than statistical precision.

Other states address the small numbers problem by reporting only the counts, and not calculating rates or averages. The rationale apparently is to
remind the user of the lack of precision in the data, but sophisticated users can figure out the denominator and calculate the rates themselves. Less sophisticated users run the risk of improper calculations or even comparing x’s without regard for differences in the n’s.

Such rules are clearly justified when applied to survey data and suppression is based on the sample size, n, in a particular category. More typically, however, these rules are used to suppress the results of infrequent events, x (deaths by cause or in precisely defined demographic groups, notifiable diseases, and so on), regardless of the size of the population, n, that generated them. Because Var (p) = p/n, rates derived from small numerators can be precise as long as the denominator is large. Suppressing the specific count only adds imprecision.

Perhaps more appropriately, some states address the small numbers problem by calculating confidence intervals. Depending on how the data were generated, different formulae for confidence intervals are available. The documentation for Washington’s VistaPH system (Washington State Department of Health, 2001) includes a good discussion of the appropriate use of confidence intervals and an introduction to formulae for their calculation.

For survey data, confidence intervals are based on sampling theory, and their interpretation is relatively straightforward. If 125 individuals in a sample of 500 smoke tobacco, the proportion of smokers and its exact Binomial 95 percent confidence interval would be 0.25 (.213, .290). A confidence interval calculated in this way will include the true proportion 95 percent of the times it is repeated.
The confidence interval can be interpreted as a range of values that we are reasonably confident contains the true (population) proportion.

The interpretation of confidence intervals for case reports is somewhat more complex. Some argue, for instance, if there were 10 deaths in a population of 1,000 last year the death rate was simply 1 percent. Alternatively, one could view the 10 deaths as the result of a stochastic process in which everyone had a different but unknown chance of dying. In this interpretation, 1 percent is simply a good estimate of the average probability of death, and the exact Poisson confidence interval (0.0048, 0.0184) gives the user an estimate of how precise it is.

A facility to calculate confidence intervals can be an option in web dissemination software or it can be automatic. A fully interactive data dissemination system, for instance, might even call attention to results with relatively large confidence intervals by changing fonts, use of bold or italic, or even flashing results. Such a system would, of course, need rules to determine when results were treated in this way. Statistically sophisticated users might find such techniques undesirable, but others might welcome them, so perhaps they could operate differently for different users.

Confidence intervals are only one use of statistical theory to help users deal with the problems of small numbers. Another alternative is to build statistical hypothesis tests for common questions into the web dissemination system. Some web-based dissemination systems, for instance, allows users to perform a $c^2$ test for trend to determine whether rates have been increased or decreased significantly over time. Some systems also allow users to perform a
c² test on survey data to determine whether there is a difference in rates between
groups. Such tests are based on implicit null and alternative hypotheses,
however, which may not be the correct ones for a given public health question.

Washington’s EpiQMS system uses a group of sophisticated statistical
techniques such as hierarchical Bayesian modeling and geographic smoothing
methods to deal with small number problems (see for example, Devine and
Louis, 1994; Shen and Louis, 1999 & 2000). An alternative model was used in
the preparation of the *Atlas of United States Mortality* (Pickle et al, 1996). In
such models, the estimated rate for a given area is based on the data for the
target area plus that for nearby areas. Depending on the variability in the data
for the target area and the desired degree of smoothing, more or less weight is
put on the nearby area data. Spatial models of this sort depend on the
assumption that geographically proximate areas have similar health outcomes,
and on this basis “borrow strength” to overcome the limitations of small numbers.
Alternatively, one could assume that non-geographic factors such as
socioeconomic status are more appropriate, and build regression models to
estimate local area rates. These and other statistical techniques for model-
based “small area estimation” (see, for instance, NRC 2000a and 2000b) are well
developed, but have only rarely been used for public health data.

Statistical models of this sort, it must be acknowledged, are better for
some purposes than for others. Because they assume relationships in the data,
such as that neighboring or similar jurisdictions have similar rates, they are not
good for looking for outliers or differences between adjacent or similar
communities. Depending on the degree of smoothing and the statistical model,
differences of this sort will be minimized. Disease clusters below a certain size will be smoothed away. These techniques can be very useful, however, in seeing the “big picture” in geographical and other patterns.

Some public health statisticians are wary about the acceptance of such complex statistical models by less-sophisticated users. EpiQMS’s developer Richard Hoskins reports that public acceptance has not been much of a problem. The software has capacity for built-in training modules, which users rely on heavily. He also notes that the community understands the basic idea that actual numbers in a given year may not be the best rate estimate, and that there is a need for the statisticians to make better estimates.

The existing web dissemination systems for public health data exhibit different attitudes about the user’s responsibility for small numbers problems. One health department official said that, to some extent, this is the user’s problem, but the web dissemination system can help by providing confidence intervals, context sensitive comments, and so on. Other states are more paternalistic, with data suppression rules and automatically calculated confidence intervals.

The developers of Tennessee’s system decided to “democratize” the data, even though this meant that some people might misuse it. They rejected the “rule of 5” for philosophical reasons. They felt that if they were seen as suppressing data it would reduce the level of trust and hurt the department’s image. The state and the center want to facilitate the use of data, and regard it as the user’s problem if they make a mistake, so the “custom query” system gives users the numbers no matter how small they are, except for AIDS deaths
(for confidentiality reasons). They warn users regarding small numbers and also recommend calculating three or five year averages. Sample data, such as from the BRFSS survey, are only reported at regional levels, but they do report small numbers. The system provides confidence intervals as well as $c^2$ tests for survey data.

An alternative approach is to offer training on the state's web dissemination system and in the use of public health data more generally. As part of the roll-out of its new web data system, for instance, Minnesota is doing regional training for local health officials on how to do health assessment. In this training, they advise people to combine data from three to five years to get reliable rates and averages. Some counties in Minnesota are so small, however, that even this doesn’t work. If so, they suggest monitoring individual cases rather than making firm conclusions.

Conclusions and recommendations

The primary statistical concern with web-based dissemination systems is the potential loss of statistical reliability due to small numbers. This is a concern in all statistical publications, but it is more acute in web-based systems because of their focus on presenting data for small geographical areas and other small groups of individuals.

There are a number of statistical techniques that interactive data dissemination systems can use to deal with the lack of reliability resulting from small cell sizes. Aggregation approaches can help, but information is lost. Small cells can be suppressed, but even more information is lost. (The best rationale
for numerator-based data suppression is confidentiality protection, not statistical reliability.) In general, approaches that use statistical approaches to quantify the uncertainty (such as confidence intervals and the use of $c^2$ tests), or smoothing or small area model-based estimation should be preferred to options that suppress data or give counts but not rates.

It must be recognized that the development of appropriate statistical methods involves a research effort that will take time and financial resources. Training people to understand and use these methods will also be resource intensive.

**Small numbers and confidentiality concerns**

Beyond purely statistical considerations, small numbers in public health data also present concerns about the confidentiality of personal and group information. Confidentiality refers to the ability or inability to identify the individuals represented in a database and use information in the database to discover their characteristics that would otherwise not be known. Although it typically refers to individuals, there are concerns about the unfair characterization of small groups and obtaining information about entities such as schools, physician groups, hospitals, and so on.

Confidentiality protection is important, first of all, because of ethical obligations of health officials to the public. It is also important in maintaining the willingness of the public and health care providers to participate in data-gathering activities that are in the public interest. Although confidentiality concerns are like
statistical reliability in the sense that the problems are greatest when the number of individuals or cases is small, the solutions can be somewhat different.

Two different types of confidentiality concerns present themselves and call for different solutions. The confidentiality issue most common in health care settings is the protection of individual medical and administrative records. Records that are used for patient care and program administration are commonly kept and transmitted in electronic form. State and federal law, including the new HIPAA regulations, protect the confidentiality of the personal health information in these records, and such records are generally not directly available to or through public health data systems (NCHS, undated). Confidentiality protection for these data is mostly focused on maintaining the security of information in databases and during transmission. Information about individuals generally only becomes available to public health authorities if they have notifiable diseases or upon death.

Government agencies increasingly make individual-level health data available to the public for research purposes. Such microdata could be the product of a population-based survey or a sample from medical or administrative records from which individual identifiers have been removed. Because the remaining characteristics may be sufficient to identify individuals, microdata are sometimes altered statistically to make it impossible to identify individuals. Continuous variables can be categorized, top-coded (every observation higher than some value $a$ is changed to $a$), or bottom-coded. Values can be exchanged between individuals, or random noise added in a way that overall summary statistics remain unchanged. Statistical techniques for protecting confidentiality
in microdata are discussed by Doyle et al. (2001) and Sweeney (1997, 1998). The confidentiality of individual-level data released for research purposes is also protected through human subjects review boards at the institutions providing the data and doing the research.

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. Being in a state’s HIV registry, for instance, could lead to stigma and discrimination, and this is also true, to a lesser extent, for sexually transmitted diseases. Recognizing someone in a public database can sometimes allow others to determine that person’s characteristics. For example, suppose mortality data are tabulated by cause of death and income in small areas, and there is only one case of lung cancer. If the identity of the person with lung cancer in the community is known, then his income will also be known. Similarly, if only one death has occurred in 50 to 60 year old men in a small town, someone might infer the cause of death from tabulated data. 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. In Washington, for instance, most of the diversity is in the western part of the state, and one case of AIDS in rural, eastern Lincoln County would be “knowable.”

Many state health data centers do not understand the HIPAA regulations which are scheduled to go into effect in 2003 and are concerned that they would restrict web dissemination of public health data. The regulations, however, are
intended to protect individual medical information, as such the regulations are focused on health care providers rather than statistical agencies. They generally do not apply to dissemination of data by public health agencies (except where health departments are themselves health care providers). HIPAA allows for health departments to collect surveillance and other data needed for public health purposes and publish them as necessary. This does not absolve health agencies from the responsibility of maintaining confidentiality, however, and web dissemination systems must take confidentiality seriously (Gostin, 2001).

No one interviewed for this report was aware of any actual problems with the unintentional release of personal data through web systems. In one state, in fact, making data available on the web has increased the level of discussion about privacy and confidentiality issues as well as the confidence in health officials and may lead to more data being available in the future.

The specific measures currently used to protect confidentiality vary substantially within and across states. The measures usually follow the rules set by data owners and are generally the same as for printed publications, so rules can vary by database. Some states treat information on the death certificate as confidential. In Washington there is no legal restriction on death certificate data, but internal regulations prohibit certain degrees of disclosure.

Minnesota, for instance, has a behavioral risk factor survey of every high school student in the state every year and will make the results available on their site. The data will be available at the county level for the public and it could be available by the individual school, or even classroom, through the state’s intranet to school officials only. In Washington’s VistaPH system anyone can “register”
data to which they then have access through the system. Data owners can also give access to other users.

One common approach to confidentiality protection in public health web dissemination systems is suppression of “small” cells in tabular reports. The “rule of 5” and similar procedures, discussed above, were in fact instituted in order to maintain confidentiality. These procedures usually work by preventing the release of information when there are fewer than \( x \) individuals in a given category. Complementary categories must also be suppressed to avoid discovery of the number of cases by subtraction. 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.

Suppression rules generally work well in protecting identity but cannot prevent someone trying to uncover certain characteristics. Suppose, for instance, characteristics of people in Town A, the smallest community in a hypothetical state, can be obtained by first tabulating data for City B (the largest in the state) and Town A together, then tabulating data for City B alone, and then obtaining the information for Town A by subtraction. A similar example could be constructed using 10 and 9 years of data. In order to avoid this type of “data snooping,” some states restrict geographical combinations for sensitive data sets. Similarly, there are algorithms to minimize the number of complementary cells that must be suppressed, but they do not guarantee non-identifiability (Federal Committee on Statistical Methodology, 1994).
What is considered “small” varies across states and across databases within states, mainly adopting the rules in individual data systems for published data. The definition may also differ from what is considered small in statistical terms, as discussed above. Garland Land has studied the situations in which personal information might be inadvertently released, as well as the likely efficacy. The following examples and discussion are based on his work (Land, undated).

The example above is a numerator confidentiality rule; data are not reported if a table cell is less than a certain number, usually 6 or fewer. The assumption behind this rule is that there are a limited number of persons with the same demographic characteristics as those in the cell, and one might associate the event in question – say a death due to AIDS – with anyone in that category – for example being Black, male, and aged 25-34. Using this rule, however, a table showing one Black male aged 25-34 AIDS death would not be published even though there may be 500 Black males aged 25-34 deaths in the population. The problem with a numerator confidentiality rule is that a small cell does not imply there are few people with the characteristics in question.

A numerator suppression rule would only allow for the margins of Table 2 to be published since the male row has one cell with 1 case and the female row has at most 2 cases per cell. Even if the rule were based on 3 cases per cell, the male row details could not be published because the “other” cell has only 1 case.
Table 2. Small numbers and confidential concerns: numerator rule examples

<table>
<thead>
<tr>
<th>AIDS Deaths in County X</th>
<th>White</th>
<th>Black</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>5/2000</td>
<td>4/200</td>
<td>1/2</td>
<td>10/2202</td>
</tr>
<tr>
<td>Female</td>
<td>2/2100</td>
<td>1/190</td>
<td>0/1</td>
<td>3/2291</td>
</tr>
<tr>
<td>Total</td>
<td>7/4100</td>
<td>5/390</td>
<td>1/3</td>
<td>13/4493</td>
</tr>
</tbody>
</table>

An alternative rule is based on population denominators; data are not reported on a population less than a certain size, such as 100,000. The assumption is that there are a limited number of persons with any given set of characteristics in a small population. With this rule, however, Table 3 could not be published.

Table 3. Small numbers and confidential concerns: denominator rule examples

<table>
<thead>
<tr>
<th>Total deaths in a county of 90,000 population</th>
<th>White</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>350</td>
<td>150</td>
</tr>
<tr>
<td>Female</td>
<td>300</td>
<td>100</td>
</tr>
</tbody>
</table>

To avoid such problems, Missouri uses a numerator and event denominator rule; a table is not reported if a table cell subtracted from the number of total events in the same data file for the same characteristics yields a small number (e.g. less than 10). For example, a cell with one Black female aged 25-34 AIDS death would be published if there were 15 Black 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 order to compare these rules, consider the examples in Table 4. A numerator rule would allow release depending on whether the number of AIDS
deaths in Black females aged 25-44 were fewer than 6 or not, regardless of the total number of deaths. The Missouri rule is based on the difference between total and AIDS deaths in this demographic group being fewer than 10. Land argues that when there is one AIDS death among 100, it is not likely that someone could identify which person had AIDS, so the numerator rule is unduly restrictive and the Missouri rule more appropriate. In the other direction, in the case where 6 of 7 deaths in this demographic group were due to AIDS, the numerator rule would inappropriately allow this information to be published.

Table 4. Comparison of numerator and event denominator rules

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

A similar case can be made against denominator rules. Consider the examples in Table 5. A denominator rule would make no distinction between one AIDS death out of 15 in the demographic group and one AIDS death out of two. Rather, the decision to release would be different if the base population were 95 or 105,000. The Missouri rule, on the other hand, focuses on whether the difference between total deaths and AIDS deaths in this demographic group is less than 10 would distinguish between one of 15 and one of two deaths, rather than on the denominator. In this example, the Missouri rule does seem to be more reasonable.
Table 5. Comparison of denominator and event denominator rules

<table>
<thead>
<tr>
<th>A: Population</th>
<th>B: AIDS deaths, Black females aged 25-44</th>
<th>C: Total deaths, Black females aged 25-44</th>
<th>Release under denominator rule: A&lt;100,000</th>
<th>Release under Missouri rule: C-B&lt;10</th>
</tr>
</thead>
<tbody>
<tr>
<td>95,000</td>
<td>1</td>
<td>15</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>95,000</td>
<td>1</td>
<td>2</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>105,000</td>
<td>1</td>
<td>15</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>105,000</td>
<td>1</td>
<td>2</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Other approaches are, of course, possible and are being actively considered. Following up on work in other states, for instance, Massachusetts is considering a different rule that also combines numerator and denominator information, as summarized in Table 6 (Cohen, 2001).

Table 6. Proposed aggregate data release guidelines

<table>
<thead>
<tr>
<th>Denominator (D)</th>
<th>Numerator (N)</th>
<th>Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 29</td>
<td>Any value &lt; D</td>
<td>Release</td>
</tr>
<tr>
<td>10-29</td>
<td>0 or D-N &gt; 4</td>
<td>Release</td>
</tr>
<tr>
<td>10-29</td>
<td>0 &lt; D-N &lt; 5</td>
<td>Do not release numerator (or complementary cells that allow for calculation of numerator)</td>
</tr>
<tr>
<td>&lt; 10</td>
<td>0</td>
<td>Release</td>
</tr>
<tr>
<td>&lt; 10</td>
<td>1-9</td>
<td>Do not release numerator (or complementary cells that allow for calculation of numerator)</td>
</tr>
<tr>
<td>= N, where D&gt; 9</td>
<td>= D, where N &gt; 9</td>
<td>Review for release</td>
</tr>
</tbody>
</table>

Data suppression rules and similar approaches to confidentiality protection in public health data systems tend to be ad hoc and generally not based in any kind of theory. Although web dissemination systems are intrinsically less
[secure] than published tables (see the Town A & City B and related examples above), public health web systems have usually adopted the rules applied to existing paper publications.

Although suppression rules of various sorts are probably the most common current approach to confidentiality protection in web-based dissemination systems, there are alternatives in place. Methods have been developed to recode, round, or perturb the data to reduce the risk of disclosure and preserve the utility of the results (Federal Committee on Statistical Methodology, 1994; Fienberg et al., 1998; Duncan, 2001).

Another alternative is to allow greater access to the database to “trusted” or “registered” users and use passwords in their web dissemination system to manage this. Health department employees and data owners, for instance, would have more access than local health officials or members of the public. Washington’s EpiQMS system, for example, currently makes data available on a “need to know” basis, which varies by dataset and geographical detail. States can also require data use agreements before trusted users are given access to data, and in this way have more control than in the past, when users were simply given a data tape. Such an approach does, however, raise questions of who is trusted, who has a need to know and ultimately issues of openness in government. Is it better to have two systems, one for health department personnel and researchers and another for less sophisticated users, than one with restricted access?

A less common but possibly important alternative is the use of smoothing algorithms and other statistical modeling techniques. Although developed to deal
with statistical reliability (see the previous section), these algorithms in effect replace the original data with a less variable estimate that does not represent any individual’s personal data. As a result, publishing smoothed values has the unintentional but desirable side effect of protecting confidentiality. Washington’s EpiQMS, for instance, also uses Bayesian smoothing to give estimates of rates for small areas without revealing specific numbers that could identify individuals.

Conclusions and recommendations

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. The definition of “small” varies by state and often by dataset. This approach often results in a substantial loss of information and utility.

Statisticians in a number of state health data centers have recently reconsidered data suppression guidelines currently in use and have developed creative and thoughtful new approaches, as indicated above. Their analyses, however, have not been guided by theory or statistical and ethical principles, and have not taken account of extensive research on these issues and development of new methods that has taken place in the last two decades. Government and academic statisticians, largely outside of public health, have developed a variety of “perturbation” methods such as “data swapping” and “controlled rounding” that can limit disclosure risk while maximizing information available to the user (Federal Committee on Statistical Methodology, 1994; Fienberg et al., 1998; Doyle et al., 2001; Duncan et al., 2001). The Census Bureau has developed a
“confidentiality edit” to prevent the disclosure of personal data in tabular presentations (Federal Committee on Statistical Methodology, 1994). Zaslavsky and Horton (1998) formulate the disclosure problem as a statistical decision problem that explicitly balances the loss associated with the possibility of disclosure and loss associated with non-publication of data. Such theory-based and principled approaches should be encouraged.

The confidentiality of data in geographical information systems (GIS) represented in maps has not been well explored. Displays that represent cases of a particular disease as dots on a map according to their place of residence, for instance, can disclose the identity if the scale is sufficiently fine and the population density is low. Random spatial perturbation has been proposed to solve this problem, but has the potential drawback of seeming to identify individuals as cases when they actually are not. As the use of maps for data display and analysis increases, research and development of confidentiality protection processes are needed.

**Concept validity and data standards**

By bringing together data from a variety of sources and making it available to a range of sophisticated and inexperienced users, the development of web-based dissemination systems for public health data brings to the surface important statistical issues in data validity and standards. Some system designers would like to provide users with a variety of different approaches to calculations, but to ensure through training and built-in help screens and documentation that they understand the concepts. Other designers focus on
users with no training in statistics or epidemiology, and build appropriate choices into the system software.

Data can be analyzed in different ways, and in some cases must be for purposes of consistency. The SIDS rate, for instance, can be calculated using births in the year as the denominator (the demographic approach), or by linking individual birth and death records (the epidemiologic approach). Both are valid, but one may be preferable to the other depending on the comparisons that will be made. Incidence and prevalence rates, as well as cause specific-mortality rates, can be presented in crude form or standardized to a common population base to clarify comparisons across time and geographical units. Whether standardization is done at all, and if so what standard population is used, whether a direct or an indirect approach is used, and the choice of age groups are all matters of judgment. Health services researchers frequently “risk adjust” hospital outcomes data to reflect different patient populations. How to do so, however, is a matter of judgment and of data availability.

Some of these issues can be addressed through data standards. Such standards might address the handling of missing data, methods for age adjustment (e.g. minimum numbers for direct adjustment), CDC surveillance case definitions, and race and ethnicity definitions. Data standards might also specify common population denominators for all rates in the system. National documents and agencies can help guide the choice of data standards. These include Healthy People 2010, and especially the Leading Health Indicators, HRSA’s Community Health Status Indicators, CDC’s Behavioral Risk Factor
Surveillance System (BRFSS), and the health care surveys and data systems maintained by the Agency for Healthcare Research and Quality (AHRQ).

MassCHIP, for instance, builds data standards and guidelines into the system through meta-data. Minnesota is planning to incorporate the emerging National Electronic Disease Surveillance System (NEDSS) standards – developed by CDC for surveillance data systems (CDC, undated) – into its state system. The NEDSS standards currently focus on details of data transmission and other technical issues rather than statistical standards, as described in the previous paragraph, but in the future NEDSS standards might expand to include more statistical issues.

Ultimately, the development of a web-based dissemination system drawing on data from various parts of a state health department and other agencies can provide an opportunity to develop or revise necessary standards. While helpful, however, data standards alone are unlikely to ensure that data are used properly. Indeed, some system designers seem concerned that providing too much data to non-sophisticated users might lead to misuse or misinterpretation.

As Sandra Putnam noted, members of the local health councils are generally not trained in health assessment; they consist of mayors, police chiefs, physicians and nurses. The University of Tennessee, Knoxville therefore has offered training in the use of the HIT system. They try to teach people about age adjustment, combining data over three to five years, but had to spend a considerable amount of time simply helping people use the web itself. In
Washington, an independent epidemiologist-demographer consultant uses VistaPH to help counties with less sophistication prepare health profiles. David Solet in Washington notes that users need sophistication in designing assessment studies and interpreting the results, not in operating the system. Much of the training for VistaPH has focused on basic epidemiology and using Excel to make graphs. Washington has also posted “Assessment guidelines” on the web (Washington State Department of Health, 2001). Richard Hoskins notes, similarly, the issue is not the technology but being sure that people know how to use it properly. Hoskins therefore does training on proper ways to do disease mapping, not just on how to use the EpiQMS to do it.

One approach to dealing with users who are not trained in public health assessment is to develop standard reports for the web-based data dissemination system. MassCHIP, for instance, has “Instant Topics” reports for Healthy People 2010, Healthy Start, minority health, and others topics. Standard reports of this type not only make the system easier to use, but also ensure that users employ appropriate and comparable variables for their community health profiles. Users might also take these reports as models for other topics. Massachusetts is also developing “wizards” to help occasional users.

Some states use a peer education approach. Minnesota, for instance, has an epidemiology users group that meets regularly. This group, which includes epidemiologists from a variety of program areas, has helped the state adopt the 2000 standard population and deal with other data policy issues. The group has also given guidance regarding the development of the state’s web-based dissemination system and its members train one another in the system’s use.
A number of the state web-based dissemination systems are currently being used as part of the curriculum in schools of public health and other academic institutions. In addition to training future users in the use of the system and in public health assessment techniques more generally, presentations to colleges and universities help to raise awareness about the system itself.

Most users, however, will not have the opportunity for such training. Web-based dissemination systems, therefore, incorporate various approaches to training and documentation. These include on-line tutorials, context-sensitive help screens, and help desks. Web-based dissemination systems also point users to external training material at the National Center for Health Statistics (NCHS, 1999), the Centers for Disease Control and Prevention, professional organizations and universities.

Tennessee’s HIT system does age adjustment, and its associated educational efforts focus on when and why rather than how. The system also provides a variety of charts, as well as user-defined comparisons and ranking tools, which are seen as a way to lead users to proper analyses. The HIT system does not use more sophisticated models because its developers feel that the system’s users are more comfortable with “real” counts and rates and are not sure how the users would interpret the results. As is common in other states, Tennessee users often want only the data for their own area, so the system does not merge data from adjacent areas.
Conclusions and recommendations

Ever since computers were introduced, statisticians have been concerned that the availability of data and statistical software would lead untrained users to make mistakes. While this is probably true to some extent, restricting access to data and software is not likely to succeed in public health. Introduction of interactive web-based dissemination systems, on the other hand, should be seen as an important opportunity to develop and extend data standards in public health data systems.

Educational efforts building on the dissemination software itself, as well as in more traditional settings, are likely to be more effective in reducing improper use of data than restricting access. For many users, such training will need to include content on using public health data, not just on using web-based systems. The development of standard reports for web-based systems can be an effective means for disseminating data standards.

Data validation

Before public health data are published in printed form, subject matter and statistical experts review the tables and charts to ensure that errors or inconsistencies in the data are found and corrected. When web-based dissemination systems create analyses that have not gone through this process, embarrassing errors can occur. This is especially true when web-based systems allow for more geographical detail than is otherwise published. Suppose, for instance, that data are available by Zip code. Transposition or geocoding errors might assign a small number of cases from a large city to a rural area with a
similar Zip code. Errors of this sort would not be noticed in state-level analyses, or in the large city. In areas with small population and few deaths, however, the addition of one or two miscoded cases would, in relative terms, be a major error.

Errors in web-based data dissemination systems are of two types. First, there are errors in the raw data and in any measures that are based on them. Second, data systems themselves can introduce errors in processing or statistical analysis. Validation approaches can address one or both of these sources. Errors occur because data systems make at least three transformations: (1) from individual records to statistical variables for different geographical areas, demographic groups, and so on (2) from variables as originally recorded to recoded measures that are more comparable across datasets and more suitable for data analysis and (3) from one data definition to another to account for differences in data standards from year to year in the same datasets. Processing of missing and unknown values can also lead to problems.

State health data centers have taken a variety of approaches to data validation. MassCHIP, for instance, strives for “100% validation.” Every time a small change is made in the system, its developers test for unintended changes elsewhere. This requires many hours of trying the system on real data, comparing to previous reports (published and unpublished), and looking for suspicious results. Substance experts are also involved in validation efforts. In Washington, there are three to four people in the department with their own programs, and they compare their results “all the time” to make sure that they get the same results as VistaPH. They use a “re-extraction” process that looks back
over 20 years of data. Similarly, before the Tennessee system is opened to the public, it is checked against every available printed report (published or in-
house). Better documentation of what is already done would be useful to other states, whether they currently have web-based dissemination systems or are developing them.

Another approach is to use the same dataset for every purpose. In Washington, for instance, the state health data center releases the same data to VistaPH and EpiQMS as to NCHS and others, so it benefits from multiple internal edits. Because there are so many users, many people look at these data and urge them to make corrections when needed. Validation efforts of this sort, however, can only detect errors due to processing or statistical calculations, since every system has the same raw data.

**Conclusions and recommendations**

No statistical techniques can guarantee that there will be no errors in web-based data systems. Careful and constant checking of both the data and the dissemination system, as well as a policy of releasing the same data files to all users, however, can substantially reduce the likelihood of errors. Methods for validation should be documented and shared among states.

The development of web-based dissemination systems is an opportunity to implement data standards rather than a problem to be solved. Efforts to check the validity of the data for web dissemination purposes may actually improve overall data quality in state public health data systems.
CONCLUSIONS

Web-based systems offer state health data centers an important opportunity to disseminate data to public health professionals, local government officials, and community leaders, and in the process raise the profile of health issues and involve more people in community-level decision making. Web-based dissemination systems, because they require that multiple data systems be put into a common format, present opportunities to disseminate appropriate data standards and thereby unify state data systems. The work required to validate the data in the systems can also result in better overall quality data in the state system.

The primary statistical concerns with web-based dissemination systems relate to small numbers. In particular, data for small population groups can be lacking in statistical reliability, and also can have the potential for releasing confidential information about individuals. These concerns are present in all statistical publications but are more acute in web-based systems because of their focus on presenting data for small geographical areas. Data suppression can resolve some of these problems but results in a significant loss of information. Aggregation, whether advised or automatic, is preferable but also results in a loss of information. Formal statistical methods (confidence intervals, significance tests, geographical and model-based smoothing, and other methods for small area statistics) allow the maximum amount of information to be disseminated while at the same time honestly communicating to users about statistical reliability and protecting the confidentiality of individual health data. While
models of this sort exist in other areas, their application in public health has been limited and should be further explored.

The further development and use of web-based data dissemination systems will depend on a good understanding of the systems' users and their needs. System designers will have to balance between enabling users and protecting them from themselves. Systems will also have to develop ways to train users; not only in how to use the systems themselves, but also on statistical issues in general and the use of public health data.

Research to develop and implement new statistical methods, and to better understand and address users' needs, is a major investment. Most states do not have the resources to do this on their own. Federal agencies, in particular through CDC’s Assessment Initiative, could help by enabling states to share information with one another, and by supporting research on the use of new statistical methods and on data system users.
REFERENCES


Health Resources and Services Administration (HRSA), 2000. Community Health Status Indicators Project. [www.communityhealth.hrsa.gov]


Appendix A: Individuals who were Interviewed or Commented on the First Draft and State Websites

**Massachusetts**
Daniel Friedman, Assistant Commissioner, Bureau of Health Statistics, Research and Evaluation

Marlene Anderka, Director, Office of Statistics and Evaluation

MassCHIP: Massachusetts Community Health Information Profile ([masschip.state.ma.us](http://masschip.state.ma.us))

**Minnesota**
John Oswald, Director, Center for Health Statistics

Peter Rode

Steve Ring

Richard Fong

**Missouri**
Garland Land, Director, Center for Health Information Management & Evaluation

Norma Helmig, Chief, Bureau of Health Resources Statistics

Eduardo Simoes, State Epidemiologist

MICA: Missouri Information for Community Assessment ([www.health.state.mo.us/MICA/nojava.html](http://www.health.state.mo.us/MICA/nojava.html))

**Tennessee**
Sandra Putnam, Director, Community Health Research Group, University of Tennessee Knoxville

HIT: Health Information Tennessee ([hitspot.utk.edu](http://hitspot.utk.edu))

**Washington**
John Whitbeck, Research Manager, Center for Health Statistics

David Solet, Assistant Chief Epidemiologist

James Allen, Senior Systems Analyst

Richard Hoskins, Public Health Geographer & Senior Epidemiologist

EpiQMS ([www5.doh.wa.gov/epiqms](http://www5.doh.wa.gov/epiqms))

VistaPH ([www.doh.wa.gov/OS/Vista/HOMEPAGE.HTM](http://www.doh.wa.gov/OS/Vista/HOMEPAGE.HTM))

**Others**
Bruce Cohen
Massachusetts Department of Public Health

Daniel Goldman, System Developer Expert Health Data Programming, Inc.

John Paulson
Center for Public Health Data and Statistics
Ohio Department of Health
Appendix B: Review Meeting Participants

Harvard School of Public Health
Boston, Massachusetts
August 6 and 7, 2002

Marlene Anderka
Office of Statistics and Evaluation
Massachusetts Department of Public Health

Steve Lagakos
Department of Biostatistics
Harvard School of Public Health

Luis Paita
National Association of Health Data Organizations

Sandra Putnam
Community Health Research Group
University of Tennessee-Knoxville

Jamie Ritchey
National Association of Public Health Statistics and Information Systems

Dan Freidman
Assistant Commissioner
Massachusetts Department of Public Health

Tim Stephens
National Association of Public Health Statistics and Information Systems

Patricia Guhleman
Bureau of Health Information
Division of Health Care Financing
Wisconsin Department of Health and Family Services

Michael Stoto
RAND and Department of Biostatistics
Harvard School of Public Health

Tim Green
Division of Public Health
Surveillance & Informatics
Epidemiology Program Office
Centers for Disease Control and Prevention

Neil Thomas
East Tennessee State University

Stepahnie Haas
School of Information and Library Science
University of North Carolina

Al Zarate
National Center for Health Statistics

Ken Harris
National Center for Health Statistics,
Centers for Disease Control and Prevention

Alan Zaslavsky
Department of Health Care Policy
Harvard Medical School
Appendix C: Glossary

Complementary suppression: Suppression of cells in the same row or column of a small cell to avoid discovery of the number of cases by subtraction.

Confidence interval: A statistical interval based on a statistical sample calculated in such a way that it will include the value of the populations statistic being estimated with a certain likelihood, usually 95 percent. The confidence interval can be interpreted as a range of values that we are reasonably confident contains the true (population) [proportion]. “or parameter” - TG

Confidentiality: The ability (or inability) to identify the individuals, small groups of individuals, or other entities represented in a database and use information in the database to discover their characteristics that would otherwise not be known.

Geographic smoothing: Statistical technique to provide more reliable estimates for small areas, based on the assumption that geographically proximate areas have similar health outcomes.

Hierarchical Bayesian modeling: Statistical technique to provide more reliable estimates for small areas, based on the assumption that non-geographic factors such as socioeconomic status are related to health outcomes.

Hypothesis tests: Statistical technique to determine whether differences (between groups or over time) are due to chance.

Microdata: Individual-level data.

Sampling variability: Uncertainty in statistical estimates based on a random sample due to the sampling itself, i.e. that repetitions of the same sampling process would yield slightly different results due to random selection.

Standardization: Statistical adjustment to reflect differences (usually in the age and sex distribution) between two populations being compared.

Stochastic variability: Uncertainty in statistical estimates due to natural variability in the process being measured. For instance, even though two communities may have the same, unchanging conditions that affect mortality, and 5 cases would be expected in each community every year, the actual number in any given year could be 3 and 8, 6 and 7, and so on, simply due to chance.