

**Proposed changes to confidentiality language are below
Changes are marked.**

II. Confidentiality

There are currently three public closed claim databases: the state-level medical malpractice databases maintained by Florida and Texas, and the National Practitioner Data Bank. These databases have proven to be an important resource for legislators, insurance commissioners, academic researchers, and others who seek to understand the dynamics of the market for malpractice insurance. Academic researchers and others have examined the causes of the premium spikes that have periodically hit the malpractice insurance market, and trends in medical malpractice litigation. When such research is used to inform legislators, and provide a guide for public policy and future reforms, it has clear benefits for all involved.

For such work to be done, academic researchers and others need access to information regarding closed claims. In deciding what information to make public, it is necessary to balance the potential benefits from having such work done by independent academic researchers and others against the confidentiality interests of individual claimants and physicians.

States must make their own decisions as to how best to balance these interests, but the framework adopted by these three databases provide helpful guidance. All three make available claim-level information in electronic format, but redact certain information. All redact patient names and provider names. NPDB and Florida provide physician specialties, as well as provider code numbers, so that one can determine if a particular provider has been subject to multiple claims. The NPDB makes physician identities available to hospitals and state medical societies, but not to the general public. Texas provides information on jury verdicts and payouts; NPDB and Florida only provide information on payouts.

The *Medical Professional Liability Closed Claim Reporting Model Law* affords states significant flexibility with respect to whether, and in what form, data may be made available to the public. This section provides three broad options designed to produce data that are analytically useful while at the same time minimizing the probability that sensitive information will be disclosed. Of greatest concern to most states is what statisticians call “disclosure risk,” or the risk that the data released could enable end-users to identify individuals or entities involved in a malpractice action. These privacy interests must be weighed against the benefits of making the data public, such as enabling independent analyses by academic researchers and others, and replicating results – two hallmarks of the scientific method.

There is a continuum of available options with respect to public release, ranging from full public disclosure to strict confidentiality. Full disclosure does not protect the legitimate confidentiality interests of physicians and claimants; strict confidentiality undermines the utility of the collected data, because it precludes analysis by academic researchers and others. Ideally, states will choose an intermediate alternative – releasing enough information to allow academic researchers and others to use the data effectively, but not releasing so much as to create undue disclosure risk.

Comments on confidentiality provisions in the Guidelines for Implementation – Part DII
October 1, 2009

David A. Hyman
Richard & Marie Corman Professor of Law
Professor of Medicine
University of Illinois
217-333-0061

The alternatives presented here are:

1. Release of individual-level “anonymized” data, in which certain characteristics associated with particular individuals or entities are either scrubbed from the data or released on more general form, and
2. A combination of partial release of claim-level data to the general public, with release of additional claim-level data (still suitably anonymized, perhaps by omitting patient names altogether, and providing code numbers for providers) to researchers who are affiliated with a reputable research institution and sign a confidentiality agreement in which they agree not to name individual providers in their research. This “confidentiality agreement” approach has long been used by the federal government to provide research access to various health care databases, such as the National Hospital Discharge Survey.
3. Release of the data at levels of aggregation that minimize disclosure risk. This final alternative may substantially reduce the utility of the data.

Option 1: Release of individual-level records to the general public

Individual-level records can be released in a way that makes it unlikely, if not impossible, that individual identities can be inferred. The discussion below applies to records available to the general public. In general, demographic characteristics, such as age, should be released in general categories (such as ages 1-5, 6-10, etc). For the particular case of medical malpractice, it is important to be able to identify baby cases, and also to link medical malpractice data to other data sources (for which common age cutoffs are age 18 and age 65). So the categories might be <1, 1-5, 6-10, 11-17, 18-24, 25-29, 30-34, 35-39, etc. In addition, care should be taken to ensure that no data records correspond too closely to unique circumstances of a case, whereby an individual could combine the data with other publicly available information in such a way as to ascertain an identity with some degree of certainty. For example, a dataset containing only a single claim against a neurosurgeon for an injury occurring on a given date within a specified geographic location may allow one to easily identify the practitioner. (Of course, if extensive information regarding the claim is already in the public domain, it is unclear what additional information might result from such “matching.” Thus, a high profile malpractice case that has already received lots of media coverage is already known, and there is no material incremental disclosure risk.) The following guidelines are intended as suggestions for states that wish to preserve anonymity while releasing data in its most usable form.

- a. References to small geographic units could be suppressed, or small units should be combined into larger units. For example, for small, rural counties, the county of injury might be combined with other nearby rural counties, which are given a single combined location identifier.
- b. Exact injury, lawsuit, settlement, or trial dates might be reported giving only the month and year rather than the exact date. If states do not wish to disclose the exact day of the month for each event, a “time from injury to lawsuit” or “time from injury to settlement” variable might be provided; states wishing to disclose these variables must also disclose month and year. Without information on month and year, time trend analysis will be impossible.
- c. The specific identify of the reporting entity may be kept confidential in individual records. However, variables describing the type of reporting entity (such as insurer, self-insured, etc) may be released without significant disclosure risk if there are a sufficient number of such entities providing medical professional liability coverage in a state.
- d. Data records that specify fairly unique characteristics of events or individuals should be suppressed, or aggregated into broader categories. For example, if there are five or fewer reports involving a particular medical specialty, the state might consider combining this specialty with a similar specialty to obtain a larger number of claims, or suppressing the specialty field. For example, specialties may be aggregated into a new, more general specialty code to attain the minimum five records in an identified geographic area.
- e. It is preferable to aggregate information – as in the county or specialty examples above – or to suppress particular fields, such as a county or specialist field, than to suppress information about an entire claim. Suppressing claims entirely will distort the entire dataset even for research for which the county or specialty was not relevant or was of secondary importance.

Option 2. Partial release of claim-level data to general public; additional release to researchers.

For data fields that could result in inadvertent release of confidential information about individual patients or providers, additional detail could be provided only to researchers, subject to the researcher signing a confidentiality agreement. For example, ages, or exact dates, or identities of small counties, or medical specialties, could be aggregated, or particular data fields could be suppressed, as part of a release to the general public, but provided to researchers who request this information. Provider or insurer code numbers could be provided to researchers but not the general public, and so on.

This option could provide a compromise between those who favor broader public access to information, those who recognize the value of providing data for research purposes, and those

who are concerned with inadvertent release of data that could be traced back to a particular individual patient or provider.

A sample confidentiality form, developed by the State of Florida, is attached below.

Option 3: Release of aggregate data

The release of aggregate data largely eliminates disclosure risk, but substantially reduces the value of the information to academic researchers and others. The Federal Committee on Statistical Methodology, under the authority of the Office of Management and Budget, has developed general guidelines to preserve the confidentiality of information collected and then publicly disclosed by federal agencies. The standards can be found in Federal Committee on Statistical Methodology, Office of Management and Budget, *Statistical Policy Working Paper 22 (Revised 2005) – Report on Statistical Disclosure Limitation Methodology*. As of August 2008, this paper is available on the internet at: <http://www.fcsm.gov/working-papers/spwp22.html>.

The most common rule type governs the statistical properties of data cells in aggregate data. The most straightforward guideline is the **threshold rule**, which is simply the requirement that a minimum number of observations appear within a data cell. Obviously, a cell count of 1 possesses a high disclosure risk. For example, assume the release of a record in which exactly one medical malpractice payment was made in 2007 on behalf of a neurosurgeon practicing in a sparsely populated county. Very likely, the individual could be identified from other publicly available information, since only a single neurosurgeon may practice in a given county.

A data cell consisting of only two observations would also pose a high risk of revealing private information. Assume that two payments were made on behalf of two physicians by two different insurers, and the data are released in aggregate. In this instance, each insurer could identify the payment amount of the other insurer simply by subtracting their payment from the total.

Obviously, the more individuals that make up the aggregate figure, the safer are the identities and of each. It is not uncommon for federal agencies to release data cells consisting of as few as three observations. A threshold of five or more may be used if the data are particularly sensitive. The threshold rule is usually supplemented by additional rules that afford greater privacy protections.

For data consisting of magnitudes (income, malpractice payments, etc.), it is likely that some cells will be highly skewed toward high-end values (incomes or malpractice payments greater than \$1 million, say). Highly skewed distributions pose a high risk that an individual could identify the highest values with a reasonable degree of certainty. A cell consisting of the sum of one very large payment and several much smaller payments would itself constitute a reasonable high-end estimate of the largest value. Knowledge of the highest value case could also permit an identification of the individual associated with the case. For example, one could search court records within a county for all cases with payouts of between \$1 million and \$2 million. As such, the Committee on Statistical Methodology has urged government agencies to adopt at least some following “sensitivity rules” *in addition to any threshold criterion*.

(n,k) rule (also called the “dominance rule”) – this rule is designed to limit access to data cells in which one or two high value observations contribute a substantial portion to the overall cell total, as in the example above. The rule is violated if some number of observations (n) exceeds (k) percent of the cell total. Commonly, n is assigned a value of one or two.

P-Percent Rule (or the “p-percent estimation equivocation level”) – This rule contemplates a “coalition” of individuals (c) pooling knowledge to estimate the largest contributor to a cell total.¹ Such individuals could be physicians represented in a cell, their insurers, or plaintiff attorneys that have knowledge of cases represented in a cell. For example, if a single law firm represented two of three cases that comprise a cell total, the firm could easily identify the value of the third contributor by simply subtracting their two cases from the total.

The rule makes the rather generous assumption that, based solely on general knowledge, estimates can be made to within 100% of the true value of each observation that comprises a cell total. In cases where “general knowledge” is less reliable, the rule will afford significantly *greater* confidentiality protections.

To limit the ability of coalitions to pool information to reliably estimate the value of subcomponents of a total, the p-percent rule constrains the percent distribution across cases that make up the total. Specifically, the rule states that any estimates derived from the data should be imprecise (or not come within p percent of the actual value). The limiting case is where the second and third largest contributors to a cell pool knowledge to estimate the largest contributor.

While the mathematical derivation and proofs of the rule are somewhat complex, the rule itself is not. It simply specifies that the sum of the remaining contributors to a cell total (everyone but the three largest contributors) must be larger than p percent of the largest observation:

$$\sum_{i=c+2}^N x_i \geq \frac{p}{100} \times x_1$$

Where

$c+2$ represents all observations but the largest three;

N is the total number of observations in a data cell;

X_i = the value being tested, such as claim payment amounts; and

p represents a percentage less than 100 to be determined by the commissioner.

In practice, the rule means that anyone with knowledge of the second and third largest observations will be able to estimate the highest value only with p -percent accuracy.

¹ It has been shown mathematically that if the value of the largest contributor cannot be estimated with accuracy, then no other subcomponent of a total can be estimated.

pq rule – This rule is derived from the p-percent rule, but assumes that a potential “coalition” could have greater knowledge than assumed in the p-percent rule. That is, the pq rule assumes that estimates of true values could be made that are much more precise than “within 100% of the true value.” This rule is not in general use, nor is it recommended by the Committee on Statistical Methodology. As such, it is not further discussed here. More information can be obtained from the working paper cited above.

The parameters in each of the above rules (*c*, *p*, *n*, etc) are specified by each agency on a case-by-case basis. **Importantly, the committee recommends that the values that an agency adopts *not* be made public, since knowledge of the parameters can aid end-users in making various estimates.**

Cells that fail a test can be collapsed into other observations. For example, data at the county level can be combined with other counties or aggregated at some other higher level of geography.

The following table is derived from the *Statistical Working Paper 22*, and describes the practices of various federal agencies with respect to the public release of sensitive information.

Agency	Threshold – minimum number for each data cell	Other threshold rules
Department of Agriculture – Economic Research Service	3	(n,k) rule –No single observation can represent more than 60% of a given cell total (see explanation of the (n,k) rule above. In this case, (n,k) = (1,0.6)
Department of Agriculture – National Agricultural Statistics Service	3	(n,k) rule , the parameter values are administratively determined and vary
Department of Commerce – Bureau of Economic Analysis	N/A	p-percent rule , value of <i>p</i> is administratively determined and varies across datasets
Bureau of the Census	Threshold varies, though the most common rule is that a cell must represent a minimum of 3 individuals from separate households	p-percent rule ; value of <i>p</i> is not published Some (sampled or micro-) data is not released on a geographic unit with a population of less than 100,000; and the most detailed micro-data the population must be at least 250,000
Department of Education: National Center for Education Statistics	3	Data is matched with all publicly available data sources. If potential

Agency	Threshold – minimum number for each data cell	Other threshold rules
(NCES)		<p>matches can be narrowed down to as few as two institutions, data is not disclosed</p> <p>Values are coded in ranges (for example, income between \$50,000 – \$75,000)</p> <p>Values are top- and bottom- coded to prevent identification of outliers</p>
Department of Energy	N/A - cells with too few observations are suppressed for accuracy reasons rather than for confidentiality (suppressed when standard error > 50%)	pq rule – values of p and q are not published
National Center for Health Statistics	n=5	(n,k) rule , parameters aren't published
Department of Justice: Bureau of Justice Statistics (BJS)	n=10	The BJS does not use any of the additional rules specified above. They do take additional measures to enhance the anonymity of the data, such as publishing values in ranges
Department of Labor: Bureau of Labor Statistics	Value of n is not released to the public	(n,k) rule , parameters not published
Department of Transportation: Bureau of Transportation Statistics	No agency-wide rule; established on a case-by-case basis	No agency-wide rule; established on a case-by-case basis
Department of the Treasury: IRS, Statistics of Income Division	n=3 for data aggregated at the state level or larger geography; n=10 for data aggregated at sub-state levels	The division does not use any of the additional rule
National Science Foundation	Does not generally rely on a threshold rule	Either (n,k) rule or the p-percent rule
Social Security Administration	n=3 at state level, n=10 at county level	

Internal Policies and Procedures

If data are confidential, each department should adopt reasonable policies and procedures to limit unauthorized access to files. Most agencies with sensitive files limit access to departmental employees who have a reasonable business- or job-related purpose to do so. A sample confidentiality form is printed on the following page. Each employee with access to confidential materials should sign the form.

Sample Confidentiality Form for State Employees

Each individual granted access to the raw or “unit level” medical professional liability closed claim data collected pursuant to [enter appropriate statutory citation] must sign this confidentiality form and initial each of its provisions.

Only employees who have a job-related purpose to access the data may do so. Access to all other employees is prohibited. _____ (initial)

Description of duties related to data (to be completed by employee’s supervisor):

An individual who has signed this confidentiality agreement has no authority to grant unit level access to any other individual who has not been granted such access. _____ (initial)

All electronic copies of data must be password protected and otherwise secured against unauthorized access. This password must not be disclosed to others who have not been granted access to the data. _____ (initial)

Paper copies of data must be stored in a secure location (locked filing cabinets, etc). _____ (initial)

Data may be released to the public only in the form prescribed by applicable departmental rules, and only pursuant to written permission obtained by the director. _____ (initial)

The process by which data are prepared for public release should be documented. A copy of the computer programs used to process the data and any resulting logs shall constitute appropriate documentation. Documents shall be retained for a minimum period of five years, as should a copy of the data that was released. _____ (initial)

Any breach of security or other disclosure must be reported immediately to your section supervisor or division director. It is the duty of the supervisor to take all appropriate steps to minimize the risks associated with a security breach. _____ (initial)

If data are stored on your hard drive, the computer must be locked and password protected when it is left unattended. _____ (initial)

Any data removed from the premises in a laptop or other electronic media should be logged, and should remain secure from unauthorized access. _____ (initial)

Authorization to access the data is automatically revoked when an individual in a position granted access leaves that position _____ (initial).

Signature _____ Date _____

A signed copy of this form shall be placed in the employee's personnel file.

Sample Confidentiality Agreement for Researchers

As an example of a confidentiality form for researchers, we attach the form used by the State of Florida, for access to patient-level data on hospital admissions and outcomes. The agreement has a defined term (currently one year), but Florida generally allows extensions to allow for research that exceeds this period.