Disclosure Risk Assessment for Microdata

By
Philip M. Steel[1]

Abstract

While researchers in government and private industry have always made efforts to maintain confidentiality of survey participants, accessibility of data through the Internet, increase in the retention of individual level data, and improvements in matching techniques have made the task increasingly difficult. In addition, protection of subjects' privacy is no longer just an ethical and professional obligation; it is now often a legal obligation. The question of when data is at risk, or in violation of law, has becomes a very real issue. This paper goes through the basics of disclosure risk assessment and explores the context in which it occurs.

Keyword: confidentiality, HIPAA, record linkage, disclosure avoidance,

Introduction

From marketing to evaluating public policy to clinical studies, solid data and good statistics are an underpinning of everyday life. The data that supports these uses are part of a general proliferation of datasets that have personal data. Since the utility of statistical data lies in the aggregate, why then is there concern over confidentiality? In order to make analyses available to public scrutiny, or to move data from the collector to analyzer, the data are generally in the form of microdata, i.e. they are at the person level, but without direct identifiers. In some cases, where data is very detailed, it is possible to reattach identifiers to some or even most of the records. Even the presence of some generic variables, for example "age", in confidential data heightens the risk of a disclosure occurring. Disclosure avoidance seeks to establish best practices consistent with the statistical use of person level data.

Data holders must understand which regulations and what standards apply to their data. The first sections look at regulations in the federal data system, then the regulation for data covered by HIPAA (Health Information Portability and Accountability Act) and in NIH (National Institute of Health) funding requirements. After looking at the legal context, we examine some background issues generally not addressed in the legal mandate: different kinds of disclosure and different concepts of risk. We then look at risk assessment in practice and give a brief overview of current research. The last sections of the paper develop an example and discuss the role of external data in risk assessment.

Regulation of Government Data

Concerns about privacy intrusion by government have led to barriers for data sharing between different parts of government and specific regulation about releasing identifiable data. These regulations are often in an agency's authorizing legislation. Title 13, which authorizes the activities of the Census Bureau, includes the prohibition that the Bureau shall not "make any publication whereby the data furnished by any particular establishment or individual under this title can be identified". The National Center for Health Statistics' statute states "… such information may not be published or released in other form if the particular establishment or person supplying the information or described in it is identifiable unless such establishment or person has consented". These statutes not only direct the behavior of the agency, but also provide legal backing for the refusal to release identifiable data. Other government agencies, realizing that confidentiality protections are a necessary part of many statistical data collections, have relied on more general legislation, citing particular FOIA (Freedom of Information Act) exemptions or sections of the Privacy Act.

The recent Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) seeks to improve and formalize some data sharing mechanisms between government agencies. At the same time, it strengthens the legal protections that agencies can employ. The named agencies may collect data under the title for statistical purposes and use a standard pledge, complete with penalty for violation. The legislation expects agencies that employ CIPSEA for a data collection to maintain a

[1] This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

high standard of protection, hoping that the public will eventually recognize and associate that high standard with the CIPSEA pledge. Agency practices and various statistical techniques are described in Statistical Working Paper #22, a publication of the Federal Committee on Statistical Methodology.

Private Sector Regulation

HIPAA extends some the same principals that have been applied to government data to health care data. It regulates the handling of medical records and billing information. Much of the regulation addresses to whom and under what circumstance health information should be made available. The section of interest here is on research access to data covered by the Act. Three methods are outlined. There is the "limited" data set, where the holder of the health data may license a partially de-identified data set for research purposes, thus passing on an obligation for confidentiality. There are also two de-identification methods, the first a safe harbor for data sets with limited geographic content and without known risk. The second is a documented process wherein a qualified individual using "best practices" certifies a de-identification.

Concurrent with the debate over privacy and confidentiality there is also a debate over openness and accountability in scientific endeavor, medical science in particular. Public funding increasingly comes with an obligation to share the data produced with the general research community. NIH requires a plan and budget for publication of the data that is produced under an NIH grant, where that is feasible. The plan, which must be a part of the original proposal and funding, is now structured to account for the cost of producing public data; this should include the work required to de-identify the data. Such research is subject to review through the Internal Review Board (IRB) process. The IRB review is primarily for adherence to the Common Rule (human subjects protection) but includes related areas, among them adherence to the new NIH regulation. IRBs are directed to seek support in their institution and will likely rely heavily on the HIPAA regulatory view, even though not all such data may be subject to HIPAA.

The regulatory view is rooted in the experience of government statistics and as such is generally conservative. This conservatism may actually be of benefit, since medical data tends to be riskier in the sense of potential for tangible harm. It is also riskier in the disclosure sense; health data is inherently more identifiable. The National Human Subjects Research Protection Advisory Committee [2002] has come out with this view:

"Many studies pose minimal risk to research subjects. Some studies [in biomedical and social sciences], however, are inaccurately perceived as conveying minimal risk. In such studies, disclosure of identifiable data may present a significant risk to the subject as a result of the sensitive nature of the topic, the variety of social interactions, or possible financial or legal implications of the activity being studied. In such research, especially in the social and behavioral sciences, protecting the confidentiality of data collected from or about private individuals is often the key element in minimizing risk.

In addition to protecting research subjects from harm that might result from their participation in research, applying appropriate confidentiality protections provides other important benefits. Confidentiality protections minimize subjects' concerns over the use (or misuse) of the data. Subjects consequently provide more accurate information to investigators, thereby improving the data used in the analysis and thus the overall quality of the research. Confidentiality protections allow researchers to continue to conduct difficult research on important societal problems."

The American Statistical Association's Privacy and Confidentiality committee maintains a web resource[2] which provides an extensive survey of documentation on data regulation.

Definitions of Disclosure

In order to formulate some notion of risk, one needs to know what constitutes a disclosure. Lambert [1993] defines three types of disclosure:

1) Identity disclosure: if a third party can identify a subject or respondent from the released data.
2) Attribute disclosure: when confidential information about a data subject is revealed and can be attributed to the subject.

Identity disclosure can occur without any attribute being divulged. It reveals the participation of the subject, which may or may not be a breach of confidentiality. Conversely, an attribute disclosure

[2] http://www.amstat.org/comm/cmtepc/index.cfm

can occur in a context where identity (in the sense of inclusion in the data set) is not an issue. An example of this is a survey of businesses where large companies are included with certainty.

Concepts of Risk

In the ruling on Southern Illinoisian v Department of Public Health we get a viewpoint on the legal interpretation of risk. The judge, ruling in favor of the newspaper's suit to obtain date, type and location data from the states cancer registry, felt compelled to add a "reasonableness" standard. He states: "The phrase 'group of facts that tends to lead to the identity' must mean any group of facts that **reasonably** would tend to lead to the identity of specific persons." He is interpreting the state law governing the registry, where it defines identifiablity. The case is made even more interesting because the state had in fact brought in an expert, who demonstrated the vulnerability of the data to re-identification. The court record of that demonstration was sealed; the expert standard was not "reasonable". An appeal is being considered in this case.

In the protection of human subjects one also finds a notion of what constitutes "reasonable" risk. The common rule states "Minimal risk means that the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests." Though that standard is reasonably clear in the context of a clinical trial, its application to a risk of disclosure is more difficult. There are two separate harms to consider, the harm to privacy (which is intangible) and the harm that may result from the facts obtained by disclosure.

Risk in the disclosure avoidance literature is the likelihood that a disclosure (of a given type) will occur. The goal is to produce a calculation that approximates the probability that an identity is correctly assigned to a data record. There is seldom an independent treatment of harm or the likelihood of an attempt to create such a disclosure. In situations where data is particularly sensitive, say HIV status, the amount of risk that is tolerated is adjusted. Rather than a focus on the tradeoff between risk of harm and benefit of research, disclosure avoidance tends to focus on the tradeoff between the utility of the data and the risk of disclosure (in the sense of correct assignment). Operations research has long had frameworks for this kind of risk assessment. The Risk-Utility framework in Duncan, Keller-McNulty and Stokes [2001] is version of this.

Modeling Risk

To make a formal model of risk, a scenario for the data intruder is required. One must make assumptions about the intruder's objectives. Do they wish to disclose the record of a particular individual, the record of any individual, or the records of at least X individuals? Is the intruder expert or casual? What data resources does the individual have? Can they draw on freely available public data? Public data obtainable by fee? Or can they use proprietary data?

Once a model has been determined, one can introduce a measure of risk, i.e. the probability that the intruder can identify a record. This requires that one model two essential components, the multiplicity in the whole population with respect to the key variables of a sample record and, in most situations, the resources at the disposal of the intruder. Unfortunately, the most vulnerable records, those that are unique in the population, may appear to be no different than other records. It is a difficult problem and even good solutions may be proximate.

Another approach (also borrowed from the data security field) is to establish vulnerabilities by "red teaming" the data. That is, assign the task (or hiring an expert) of finding an actual disclosure. This is particularly profitable because it can reveal bad assumptions about the resources of a data intruder. It also introduces a practical constraint--there is certain to be a gap between what is theoretically possible to do and what people in fact can do. For a good example of an actual data intrusion see Ochoa (2001). This is the write-up of a project for an undergraduate course at MIT.

The practice of disclosure avoidance in statistical agencies relies heavily on the employment of a checklist. See FCSM [1999] for a generic version. It looks for a variety of problems, but the central element is the classification of variables into non-exclusive categories. The checklist probes for the presence of certain types of variables and whether they are being treated properly:

- Identifiers
- Unique to data set
- Contextual
- Sensitive

- Variables descriptive of survey design or source
- Key variables

Identifiers are items more or less unique to the individual and exist independent of the data set (random identifiers are ok). These must be removed for de-identification of any degree. HIPAA lists many of them specifically:

Names
ALL geography smaller than the state
All elements of dates smaller than a year (i.e. birth date, admission, discharge, death, etc.)
Device identification numbers
Phone numbers
Fax numbers
SS numbers
WEB URL's
Internet IP address numbers
E-mail addresses
Certificate/license numbers
Vehicle identifiers
Biometric identifiers (fingerprint, voice prints, retina scan, etc)
Full face photographs or comparable images
Medical record number
Health plan beneficiary
Any other unique number, characteristic or code.
Any other account numbers

The items listed have been truncated in some instances. In particular, the first item includes a provision for 3 digit zip codes, with a population threshold applied. The list also includes some quasi-identifiers. Geography is a quasi-identifier. These are items that can be combined with another variable or other variables to form a true identifier.

An item unique to the data set is something that can be excluded from disclosure analysis since it is not available elsewhere to assist in linking data, nor is itself identifying. It can be ignored unless it falls into the sensitive category as well.

Contextual variables provide links to other lists and can be particularly damaging to efforts to limit geographic detail. A poverty rate can be returned to the list from which it came, to determine a limited list of geographic possibilities, that limited list then brought back to the file and screened against the presented geography.

Sensitive variables address the overall risk framework … sensitive variables are the ones that raise the stakes. They are the variables that could be associated with harm. Knowledge of HIV status is the common example.

Variables descriptive of survey design or source often have geographic content or otherwise add detail to published variables.

The last classification is crucial; the validity of most analyses depends on the accuracy of the expert's designation of the key variable list. These are the variables that have some counterpart on data that is in the public domain (or held privately by another party, if that is a concern). Within the record linkage terminology, these are the variables available for blocking and matching.

Current research on risk measures

There has been a great deal of activity in the research community on techniques to insure confidentiality of data, and many employ some sort of measure of risk. I have selected a few to mention here. K-anonymity has been around the longest. The "per record" risk is furthest in application. The "special uniques" idea is one of the more appealing in the directness of the approach.

The most rigorous standard is k-anonymity. A data set is safe if there are at least k (usually k=3) records that are identical with respect to the set of key variables. Suppose the key variables are age (single year), sex, type of cancer, and yes/no treatment A. Then there must be k 45-year-old men with prostrate cancer receiving treatment A in order to show that combination in the data. Data either conforms to this or not. There are procedures to produce an optimally conforming data set based on the original, but one may lose detail on age, have some record entries suppressed, etc., see Sweeney [2002].

The problem here is clearly a matter of not accounting for the protection of the data inherent in sampling. You may know a priori that there are thousands of 45-year-old men with prostrate cancer receiving treatment A, but there is only 1 in the sample data. How can you account for sampling's effect on multiplicity? There are a variety of estimates, but generally if these estimates are unbiased then they have high variance; this problem is the flip side of the famously intractable species problem.

There is a model which gives good results in some applications put forward by Benedetti and Franconi

1998. Benedetti and Franconi make their estimate based on the sampling weight of the record. This risk measure is implemented in mu-ARGUS, a product of the Computational Aspects of Statistical Confidentiality project (see CASC [2004]). In this model, weights set to one gives back an approximation of the k-anonymous model, so this is less restrictive. ARGUS also has a fudge factor entered in such a fashion as to simulate the effect of differences between the key variable values of the protected data and the key variables of the attacker's file. Applicability of the model has come into question (see Rinott [2003]). Regardless, this is an improvement over the k-anonymous procedure in the prior version of ARGUS.

One interesting factor common to these efforts and also found in Lambert is the inclusion of probabilistic assignment by the attacker. If the attacker has access to the universe from which the sample is drawn and finds 3 possible records to attach, the probability of a correct assignment is calculated and added to the overall risk. This would rate a data set with a third of the records open to disclosure the same as a data set where every record has a one third probability of a correct assignment—where the attacker has no way to distinguish between correct assignments and incorrect assignments. It seems not to account for the false match rate, and it is reasonable to believe a high false match rate would deter most attackers.

Another approach is being developed by Elliot [1998], dubbed "special uniques". Here the effort is to find risky records and is based on the notion that unique sample elements that remain unique despite aggregation should score highly. It is within the same sort of framework as Willenborg and De Waal [2000]'s "fingerprints" and is computationally intensive. Like the other methods in this section, it requires a fairly substantial data set.

Evaluation of risk in practice

Absent a generally applicable measure of risk, how do we proceed? The operation of a Disclosure Review Board (analogous to the IRB) is essentially legalistic. The checklist establishes the relevant facts for a given data release, precedence is sought, and rules consistent with past determination are applied. When new situations arise, research is done and remedies are suggested. When flaws in old rulings are found, they are revised. Some of the Census Bureau's rules are in the public domain, for example the rules applied to special tabulations of the Decennial Census and the American Fact Finder Advanced Query System (Hawala, Zayatz and Rowland [2004]).

What elements should be found in a certification of de-identification? The expert should reference the framework they are working under, justify the selection of key variables, state their knowledge of current public data, show frequencies where appropriate, document the effect of any disclosure avoidance procedure applied. For more on documenting de-identification see Rasinski and Wright [2001].

An example

How does one use a key variable list? The best way to proceed is by example. Physicians (and to some degree any licensed professionals) are a difficult group to handle because of the amount of data available in the public domain. Searchsystems.net provides an extensive directory of public data. Through it you can quickly find the Nebraska Health and Human Services License Information System. Their query system will allow implicit wild cards … selecting physician and typing "a" will return all doctors with last name beginning with "A" in Nebraska. A similar system exists for Texas. You will find that the Texas data has a more complete record of specialty. License data by its nature is both public and identified. Because their data is particularly accessible, those two states will be used for the example, with the reader encouraged to go on line and examine the resources available to a data intruder.

Suppose a study (of any size) involves physicians in a particular state. Suppose also that the key variables for the dataset are age, sex, marital status, specialty, and practice size. All these items can be found elsewhere in identified data. We need to know specifically how these items are presented. Here we assume the age range for practicing physicians to be 29-68, the specialization collapsed to 10 categories, and the practice size categorized as single, small, medium and large. The first step in evaluating risk is to do a calculation relating the key variables and the population the data resides in. The cross-categorization of age, sex, marital status, specialty, and practice size has 40x2x2x10x4=6,400 cells.

The 2000 Census shows Nebraska having roughly 3,500 physicians and Texas about 50,000. If our

hypothetical data set involves physicians from Nebraska, the classification by key variables uniquely determines most physicians in that population (3,500 physicians into 6,400 cells) under any reasonable distribution of data. On the other hand, the average cell size for Texas is 7.8, and one can safely assume, again with a reasonable distribution of the data, that the majority of both records and cells are protected. They are protected in the sense that, with respect to the overlap with external files, most physicians are indistinguishable from at least 2 other physicians in the population.

What do we mean by "reasonable distribution" of data? Consider some extremes for Texas. First the worst possible distribution: 6,399 cells each with one physician and one cell with a count of 43,601. The majority of records are protected but the majority of cells are definitely not. But we know that it is impossible that 43,601 physicians in Texas are male general practitioners, age 45, married and practicing by themselves. Nearly as unlikely is the other extreme, that most cells have 8 (some have 7) and all are amply protected. The data will distribute somewhere in between, so that there are a few cells with 0,1 or 2 on the periphery and some heavy clumping in popular practice types and younger ages.

If the standard is that the data is protected from attacks which divulge the identities of a fixed number, and that number is significant fraction of the sample, then the Texas data is safe without alteration. But if the desire is to protect all the data subjects, a good average cell size is insufficient. If the data set is a census (the whole population) then risky records are easily identified. They fall into cells of size 1 and 2 in the cross-classification.

What happens if the dataset is half the population? A tenth of the population? The percentage of data flagged as unsafe goes up and at the same time the percentage of records for which this designation is correct goes down. For large samples there is a great deal of structure that can be exploited, and one of the methods in the research section may be applicable. But if the data set is a small fraction of the population and the key variables of sufficient number, these methods will fail to predict which records are particularly vulnerable.

In absence of enough data in sample to make a good estimation of the population characteristics, one can instead go about the task of constructing the population from outside sources. The population counts that I have been using come from the 2000 Census. In fact, the Census 2000 public use microdata for Nebraska provides occupation, age, sex, and marital status. This is just a 6% sample of the population, but it can provide some indication of how the data is distributed, in particular it provides good estimates of the marginal distributions of the full population. But we can do even better than that … the license data is freely available and can be extracted by means of a web spider. That is, it can provide the actual population (or pretty close to it) for several of our key variables. When the sample data is compared to the full license set, it may already be evident which cases are potential problems. At the very least, we can restrict the modeling or data exploration to the cells in problem areas. The license data doesn't provide marital status or practice size. The census provides a distribution of age sex and marital status for that population. So, with some time investment, one could put together a pretty good model. See Duddek [2004] for an example of a "data assist".

We have not tapped other possible data sources. There are a wealth of physician finder and evaluation services. It might be prudent to subscribe to one or more to determine what information is provided by them and whether their query systems allow one to examine all records. In particular, they may be able to provide a link from the doctor to the practice. In Texas, the practice has a separate license; it may be common practice to use the same address of reference on both. This would allow one to apply address matching software and acquire a certain percentage of links. Or it may be the case that you cannot find any link between physician data and practice data and wish then to reconsider the inclusion of practice size in the list of key variables.

Going through the population construction exercise is useful. Along the way you may note relevant facts, like the proportion of male to female, the distribution in to specialty categories and so on. It may turn up marginal uniques (like a 65 year old female oncologist), which require protecting. Most importantly, you are going through the preliminary steps that a data intruder must take and you may be able to form some opinion about the feasibility of the intruder scenario you are using.

Watching Data

It is necessary to keep track of the general availability of data. What can be obtained by a simple search on name and state has gone from

what appears in a phone listing to including age as a matter of course. Birth date and geo-spatial data are also available. What relevance does finding age in a people-search engine have? It indicates that birth date is so widely known that one can assume that almost every identified public file either includes it or could be extended to include it. E-detective services and commercial mail list providers also give an indication of what kinds of data are widely available.

Online public records are particularly dangerous. Property records and other record systems legally required to be open to inspection are often brought online with a query system built-in. It makes the data much more useful for a variety of purposes. For instance, rather than "who owns property x" one can acquire a listing of "all property owned by y".

There are factors at work against the general application of record linkage. Extensive systems of records are expensive to maintain. Without proper maintenance they go out of date rapidly. It is currently the domain of government and large corporate entities to create such on-going enterprises.

This discussion covers only general lists. There are a variety of publicly available data that apply to very specific populations. Professional organization membership or hospital discharge data are two examples. Familiarity with those resources requires subject matter knowledge; this is why it is essential that individuals knowledgeable of both the design and content of a survey (or sample) fill out a checklist.

Technologies to watch

Record linkage as currently practiced comes in two varieties: rules based or (Fellegi-Sunter) probabilistic matching. Both require expertise to do well, and are invariably accompanied by "data cleansing" and unduplication operations. There now are a large number of companies who specialize in this field. Most record linkage operations are proprietary and did not advertise or distribute their software. However, recently several open source projects have surfaced on the Internet. FEBRL, Freely Extensible Biomedical Record Linkage, has a well-structured and documented set of code. SFnet describes the system as doing "data standardization (segmentation and cleaning) and probabilistic record linkage ("fuzzy" matching) of one or more files or data sources which do not share a unique record key or identifier." There is also an open source SAS based system developed by the Substance Abuse And Mental Health Services Administration (SAMHSA). Both are oriented toward medical records, but could be adapted for other kinds of data. Ease of use and resource issues deter use of record linkage for attacks on statistical data. Technical advances eat away at that deterrence, e.g. "big match" (Yancey [2002]) can overcome problems with blocking criteria. Classifiers (see Hastie [2001]) may replace probabilistic record linkage and uncouple linkage and data cleansing operations. Because un-duplication and longitudinal association both have high payoff for business applications, record linkage will likely be integrated into business management software as soon as it is feasible to do so.

On line analytic processing (OLAP) is a powerful technology still in its infancy. Considerable benefit can be realized by allowing remote queries. With such a system, an emergency response team can query a medical archive on treatment outcome for services while at an accident scene, an ability that could save lives. Such a system is being tested in Canada already. But it is extremely difficult to guard such systems against intruders, particularly those using tracker queries (see Wang et al [2003]). Because both the benefits and the risks are high, this will likely be the focus of legal tests.

Conclusion

The wide availability of basic personal data and data combining technologies has changed what constitutes identifiable data. Government data has long been regulated and measures to prevent re-identification have been developed for that need. More data are being regulated and the use of disclosure avoidance techniques is becoming more common. This paper has examined how risk is determined or evaluated.

What role a "reasonableness standard" have? In going through the basic mechanics of risk analysis, I have tried to point out a few areas where there are choices in the measurement of risk that have some effect on the conservativeness of the assessment. If the analysis must resort to a model of the population with respect to the key variables, is there a reason to believe that an attacker can do better and actually produce the population or some part of it? This is the question that the practitioners must answer.

Bibliography

Duddek, C. (2004), "Borrowing strength from census data to assess survey disclosure risk", to appear in the Proceedings of the American Statistical Association's Joint meeting in 2004.

CASC (2003), "µ-Argus 3.1 user's manual", p11-13, http://neon.vb.cbs.nl/casc/

Duncan, G., Keller-McNulty, S. and Stokes, L. (2001), "Disclosure risk vs data utility: the R-U confidentiality map", Technical Report LA-UR-01-6428., Statistical Sciences Group, Los Alamos, N.M.: Los Alamos National Laboratory.

Elliot, M., Skinner, C., and Dale, A. (1998), "Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk." *Research in Official Statistics*, 1(2), 1998.

FCSM (1999) "Checklist on Disclosure Potential of Proposed Data Releases", Prepared by Interagency Confidentiality and Data Access Group (An Interest Group of the Federal Committee on Statistical Methodology). http://www.fcsm.gov/committees/cdac/cdac.html

Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning*, Springer, New York.

Hawala, S., Zayatz, L., and Rowland, S. (2004)," American FactFinder: Disclosure Limitation for the Advanced Query System", Journal of Official Statistics March, Vol.20, No.1, 2004.

Lambert, D. (1993), "Measures of Disclosure Risk and Harm", *Journal of Official Statistics*, vol 9, 407-426.

National Human Subjects Research Protection Advisory Committee (2002), "Recommendations on Confidentiality and Research Data Protections" from the July 30-31,2002 meeting.

Ochoa, S., Rasmussen, J., Robson, C., and Salib,M. (2001), "Reidentification of Individuals in Chicago's Homicide Database: A Technical and Legal Study." Toward completion of coursework for MIT's Ethics and Law on the Electronic Frontier.
http://web.mit.edu/msalib/www/writings/index.htm

Polettini, S. (2003), "Some Remarks on the Individual Risk Methodology", Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, paper #18. Online: http://www.unece.org/stats/documents/2003.04.confidentiality.htm

Rasinski, K., Wright, D. (2001), "Practical Aspects of Disclosure Analysis", working paper NORC
http://www.src.uchicago.edu/prc/publications.php

Rinott, Y. (2003), "On models for statistical disclosure risk estimation.", Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, paper #16. Online: http://www.unece.org/stats/documents/2003.04.confidentiality.htm

Sweeney, L. (2002), "Achieving k-anonymity privacy protection using generalization and suppression.", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 19(50,2002; 571-588.

Wang, L., Wijesekera, D. and Jajodia, S. (2003), "OLAP Means On-line Anit-Privacy", Department of Information and Software Engineering George Mason University Technical Report Series.

Willenborg, L. and De Waal, T. (2000), *Elements of Statistical Disclosure Control*, Vol 155, Lecture Notes in Statistics, Springer-Verlag, New York.

Yancey, W. (2002), "Big Match: A Program for Extracting Probable Matches From a Large File for Record Linkage", United States Bureau of the Census, Statistical Research Division Research Report Series. Online at: http://www.census.gov/srd/papers/pdf/rrc2002-01.pdf