Journal of Privacy and Confidentiality (2011) **3**, Number 2, 83–90

Privacy and the Statistician: What Do We Need to Know to Certify Nondisclosure?

A comment on Gerald W. Gates's *How Uncertainty about Privacy and Confidentiality Is Hampering Efforts to More Effectively Use Administrative Records in Producing U.S. National Statistics*

Alan M. Zaslavsky*

Abstract.

Statisticians are being asked to determine what methods must be applied to protect the privacy of research subjects. Careless, venal, or malicious release of confidential data can be harmful to research subjects and to the entire research enterprise. Under the Health Insurance Portability and Accountability Act (HIPAA), statisticians may be asked to certify the adequacy of nondisclosure methods for confidential health data used in research. Yet despite extensive methodological research on deidentification of datasets, we lack a comprehensive framework for evaluating risk and selecting optimal strategies for protecting confidentiality with minimum impact on research. Among the factors that should be considered in such a framework are (1) the properties of the population or sample under a nondisclosure regimen, (2) the availability of external key databases, (3) the nature of the intruder, (4) the losses associated with disclosure, (5) the analyses to be supported by disclosure-protected data, and (6) the information losses due to disclosure protection. Although the present state of our analyses does not afford a ready answer to the ethical dilemma faced by the statistician asked to certify the adequacy of nondisclosure measures, this perspective does help to identify some of the social questions that must be addressed in considering how to protect privacy.

Keywords:

Nondisclosure, confidentiality, certification, HIPAA, ethics, deidentification, limited access dataset

Jerry Gates has written an excellent and very revealing discussion on the impact of a range of privacy and confidentiality concerns on the effective use of administrative data for government statistics. He writes with an insider's knowledge from his experience as the senior privacy officer at the Census Bureau, as a participant in numerous interagency committees on the topic, and as an inspiring chair of the ASA's Committee on Privacy and Confidentiality, among other contributions. I benefitted from his ability to innovate and overcome bureaucratic obstacles when he facilitated the establishment of what might have been the first (informal) extramural Census Research Data Center to allow research with confidential data by myself and a few colleagues. (Its physical manifestation was a VAX 750 computer we drove up from Suitland in 1990 to a new

*Department of Health Care Policy, Harvard Medical School, Boston, MA, mailto:zaslavsk@hcp.med.harvard.edu

home in a corner of the Boston Regional Office.)

His article is important both because of what it tells us about obstacles to sharing data within government and because similar obstacles are encountered outside of government in bringing together datasets owned by different organizations. I first comment on the form these issues take in my world of health services research and then on issues related to some of the points he raises concerning research needed to better assess the risks of data sharing (page 22 of his article).

1 Privacy and Confidentiality as Growing Concerns

Changes in technology have made issues regarding privacy and confidentiality more salient not only within the federal government, but across society as a whole. (I refer primarily to the United States, but similar issues apply elsewhere, notably in the European Union.) More data, including sensitive medical, financial, credit, and legal records, are available in readily transportable electronic forms. Access by many legitimate users creates the risk of inadvertent or malicious disclosure. At the same time, readily accessible online (publicly and privately maintained) and commercially available data products are both outlets for data that might be regarded as confidential (such as credit information or even social security identification numbers) and as sources of identifying keys that can be used to identify individuals in supposedly deidentified datasets. Thus the implicit protection of confidentiality in the past, through the sheer inaccessibility of data in paper records that are physically distributed or difficult to use, no longer offers the protection it once did.

Privacy and confidentiality have become flashpoints for public concerns over the misuse of personal data. These fears are generalized and not necessarily based on a well-informed assessment of relative risks in different areas. Many routine uses of data outside of research are only nominally voluntary and pose greater risks. For example, the cost of declining to allow your health data to be shared with an insurer is effectively a loss of health insurance; refusing to allow your financial information to be shared cuts you off from consumer credit. Routine uses of such data could also subject an individual to serious personal consequences, such as being denied insurance, losing employment due to the potential for high medical expenses, or losing employment due to a previous but irrelevant criminal conviction. Yet, paradoxically, such uses tend to be insulated from concerns about privacy because the individuals affected are not given an opportunity to grant or withhold consent and might not even be aware that their data have been accessed. Indeed, legal use of medical and commercial data might have even greater consequences than those of intrusions. Thus, efforts to secure data against unauthorized use or to prevent access by hackers might become ineffective proxies for the control that is denied to us in many areas of our lives over the substantive and legal administrative uses of data.

Research use of data, which is tightly monitored in the interests of the subject and often requires explicit and voluntary consent, can become the focus of concerns simply because it is voluntary and its benefit to the subject is not immediate. This appears to

be misdirected given the pervasiveness of relatively uncontrolled data sharing outside of research, the presumably benevolent motivations of researchers, and the numerous controls under the Common Rule and other rules on the principle of "do no harm" in research.

For example, concerns about routine electronic transmission of medical and billing data has led to a promulgation of privacy standards under the Health Insurance Portability and Accountability Act (HIPAA) by the U.S. Department of Health and Human Services after prolonged public comment and discussion. While the transmission of data for healthcare provider and insurer operations, including billing and profiling of patients, is subject only to technical standards, the content of data disseminated for research use is constrained by a series of limitations intended to prevent reidentification.

2 Effective Nondisclosure Protections in Research

Prevention of unauthorized disclosure involves a combination of controls on access and controls on content. At one extreme, a public-use dataset that is distributed without restrictions (e.g., posted on a website) must rely entirely on statistical controls; this is typical of public-use census data. Controls on content broadly constitute the territory of "statistical disclosure control" and involve procedures such as randomly perturbing data, masking or swapping some data elements, or replacing the original data with synthetic data with similar distributions but no identifiable records. Statistical disclosure control is a growing subspecialty within statistics (as reflected not only in numerous conference presentations and technical publications but the very inception of this journal). At the other extreme, a dataset that is closely held within an administrative agency (such as the IRS) must contain the full set of identifiers required for the agency to perform its functions with individuals, but would be subject to extremely strong controls on access.

Research typically uses a combination of the two types of controls. For example, many public-use datasets are made available only under data use agreements and licensing arrangements, ensuring that even anonymized data are only provided to legitimate research groups that restrict access to authorized researchers. Some agencies offer controlled access through restricted data centers where authorized users conduct analyses under the supervision of agency staff who vet each analysis for disclosure potential. Outside of government, individual researchers rarely have such elaborate access-control systems but may be asked to certify to data providers or institutional review boards that appropriate controls are in place.

The access-control side of data security only works if the basics of security are observed—maintaining computer security upgrades, using access controls properly, proper management and security of paper records, logging and auditing accesses to confidential data, and so forth. Computer security experts say that most break-ins are due to simple failures to follow basic procedures such as upgrading software and setting up proper permissions, passwords, and other access controls.

The deidentification requirements under HIPAA allow the "covered entities" (health-

care providers, insurers, and the data aggregators that serve them) to release data when certain standards have been met. Among the options mentioned in the rule are (1) removing 18 listed identifiers (including fine geographical information on patients) from the dataset, (2) creation of a limited data set under a less stringent standard but released under a data use agreement that includes access restrictions, or (3) other data release when a statistical expert certifies de-identification. This last option, which places the burden of assessing the risk of disclosure on the (undefined) statistical expert, was the occasion for organizing the panel discussion at which this discussion was presented.

HIPAA seems likely to have two unforeseen effects. First, an industry might arise to provide statistical certification of deidentification, with specialists setting their own standards and thus providing cover to data providers seeking protection for their data releases. Second, there has been a tendency for the crude description of deidentification in the HIPAA rules to become the de facto default standard for data release, at least in health care.

3 What Do We Need to Know?

Although the literature on disclosure control is extensive, rigorous specification of the nondisclosure control problem is limited by the difficulty of stating the required parameters. Gates's comments (page 22 of his article) suggest some directions for practice, policymaking, and research. Expanding on his list, I suggest the following elements that would be required to permit a rigorous analysis of the appropriate requirements for nondisclosure, along the lines suggested by the "toy" example of Zaslavsky and Horton (1998) and other researchers.

3.1 The Properties of the Population or Sample Under a Nondisclosure Regimen

The typical model of an "inferential intrusion" is to find characteristics (keys) in a target dataset that identify a specific individual in an external data source that contains specific identities of individuals, such as names or social security numbers. The properties of the population or sample we are concerned about are those that affect the existence of such uniquely identified cases ("uniques"), given other assumptions such as the availability of external databases.

If the identification is not certain then we might suppose the burden of disclosure to decline rapidly with the amount of uncertainty. However, near-uniques might permit identification of a case in the database down to a small number of possible matches, allowing the intruder to claim (misleadingly) to have identified a subject or to gain probabilistic information (such as that an individual has an elevated likelihood of having AIDS).

If the dataset is based on an anonymous sample, then population uniqueness might be of primary interest. On the other hand, for example, suppose that a parent knew that

his child had answered a survey on sexual behavior or drug use and wanted to identify the child's answers: then sample uniqueness would be relevant. Thus, a key piece of information for reidentification might be survey participation itself, which is generally known only to the participant. Some datasets, such as those based on a disease registry, represent a census of a defined population, and then population and sample uniqueness are equivalent.

There is a substantial literature on predicting the number of uniques with various populations and choices of keys; the perturbation methods used to mask the data are another input to determination of uniqueness in a masked dataset.

3.2 Availability of External Key Databases and Other Identifying Information

The potential for reidentification depends on what keys can be used, and hence on what databases are available for matching. New data sources are constantly becoming available, including some that we might not even be aware of when we do our assessment of risk, or that we might believe to be less accessible than they actually are. In some cases, anecdotal information that cannot be obtained from any systematic source might also be useful for reidetification.

3.3 The Nature and Resources of the Intruder

The characteristics of the hypothetical intruder are critical to analyzing the intruder's chances of success, yet this area remains murky, in part because there may have been few enough recognized cases of intrusion (or few enough that have been publicized) to permit generalization.

The intruder might simply want to identify a few cases in order to create embarrassment to the data-holder. Slightly paraphrasing one expert informant, "The documented cases of intrusion I know about are limited to those conducted by deidentification experts proving the importance of their work." He might want to obtain information on a specific subject (perhaps preferring a notorious one to maximize publicity, e.g., the governor, or simply a particular individual he intends to harass or victimize); in that case the existence of a small number of identifiable cases is irrelevant unless the target case is one of them. Or he might wish to obtain information on a whole list, for example to sell information to a corporation that can use it for sales, insurance underwriting, hiring decisions, etc., in which case identification of a few unusual cases might not be worth his trouble.

The skills and computational and information resources of the intruder are also crucial for distinguishing between the theoretical and practical risks of disclosure. What key databases, or information on specific target cases, are available to the intruder? Does the intruder fully understand the nature of the measures taken against disclosure? (Idiosyncratic, undeclared aspects of dataset perturbation give more protection than would be afforded if the same procedures were published as part of the data spec-

ification.) How well can he apply the algebra of Markov move bases, or solve linear programming problems?

Because of the complementary role of statistical and access-based nondisclosure protection, the ability of the intruder to penetrate access controls is also important: do we fear the insider or the outsider? Finally, the hypothetical malice and determination of the intruder affect the implications of a potential reidentification and disclosure.

4 Losses of Disclosure

When we choose a deidentification policy we implicitly or explicitly quantify the losses associated with an undesired disclosure. We might wish first to consider the nature of losses imposed on those whose information is revealed, such as legal consequences, administrative sanctions, monetary exposure, and various kinds of indirect consequences. Disclosure is only significant if sensitive information is revealed, and lags and survey errors may reduce the value of the information and hence the loss to the subject. Furthermore, the scale of disclosure affects aggregate losses of this type, which might be nearly proportional to the number affected. Conversely, the loss to the data provider or the sponsor of a research project (or to research in general) might depend much less on the scale of the disclosure, since even a handful of well-publicized cases could be damaging.

Criteria such as "minimal danger of disclosure" or "following accepted standards" are not helpful in defining costs or bounding acceptable risks.

5 The Analyses to be Supported by Disclosure-protected Data

To consider the losses associated with masking of data, we should know what analyses the data are intended for. Nondisclosure implies that some questions cannot be answered correctly and should be designed to allow the answering of desired questions and not those that infringe on confidentiality.

For tabular data we might ask which cells or margins need to be correct, and what accuracy is required. Similarly, for analyses of microdata we might ask which margins and relationships will be studied.

A substantial research effort has been devoted to developing corrections for the effects of disclosure-limiting perturbations of data. Formal model-based methods might allow us to assess the "congeniality" of particular disclosure-limitation methods and analyses better than is possible with ad hoc procedures.

6 Information Losses Due to Disclosure Protection

Two types of losses might occur due to imposing disclosure-limiting restrictions or perturbations on a dataset. First are the losses due to increased inferential variance. The nondisclosure strategy ideally would include a valid way of assessing this variance. Thus the loss can be quantified as increased variance or equivalently, as decreased effective sample size.

It is more difficult to quantify the losses due to bias (essentially the inability to do some analyses and obtain valid results). The worst outcome might be an inability to recognize when analyses using masked data are wrong!

A substantial amount of work has been done on quantifying the loss of information due to masking. More broadly, it would be desirable to have a common utility metric for comparing the "costs" of losing information to the "costs" of disclosure.

7 Conclusions

In our present state of sophisticated but still limited tools and understanding, many ad hoc solutions have been proposed, and indeed a small industry is building up around nondisclosure. Yet we don't know enough to do a solidly-based formal assessment that might guide our practice. Some methods might be preferable simply because they lend themselves better than others to assessment of at least some of the criteria. In particular, model-based methods allow a more systematic assessment of the effects of nondisclosure measures on inference.

One of our key practical challenges is to better understand what we are defending against, so that we can direct our efforts against the worst threats. At the same time, we must work to improve understanding (among policymakers and the general public) of the differences between research and the administrative and enforcement uses of data. On the technical side, we need to routinize methods so that they can be applied more widely, on the principle that statistical methods are research tools until they are embodied in readily available software.

When can we certify nondisclosure? If we are too restrictive, we hurt potentially valuable research, but if we are too permissive, we threaten harm to subjects and/or to the research enterprise. Thus, conflicting values must be balanced by experts hampered by limited time, knowledge, and expertise, and conventional standards and "off the shelf" technologies might be inadequate.

We live in a time of great confusion and great opportunities with respect to privacy and confidentiality. My hope is that the statistical community can address these challenges on a broad front to develop socially beneficial solutions. Jerry Gates's thoughtful account gives some grounds for optimism.

References

Zaslavsky, A. M. and Horton, N. J. (1998). Balancing disclosure risk against the loss of non-publication. *Journal of Official Statistics*, 14: 411–419.