SSC Annual Meeting, May 2006
Proceedings of the Survey Methods Section

RECORD LINKAGE, NONDISCLOSURE, COUNTERTERRORISM, AND STATISTICS

Michael D. Larsen[1]

ABSTRACT

Governments collect information for many purposes, including preventing terrorism. Record linkage is the process of merging databases containing common units and has many applications. Both false and missed links between records have potentially severe consequences in the context of counterterrorism. Use of record linkage methods for counterterrorism could affect other data collection efforts in terms of increasing nonresponse and decreasing data quality. Most government data collection efforts include guarantees of privacy, confidentiality, and nondisclosure. The focus of this talk is the relationship between advances in record linkage, disclosure risk, and counterterrorism efforts involving large databases.

KEY WORDS: Data fusion; Disclosure avoidance; Disclosure control; File matching; Inference control; Privacy.

RÉSUMÉ

Les gouvernements rassemblent de l'information pour plusieurs raisons, incluant l'empêchement du terrorisme. Le couplage d'enregistrements est le processus de apparier les bases de données avec les unités en commun, et il a beaucoup d'applications. Les liens faux, et les liens manquants entre les unités ont des conséquences potentiellement graves dans le contexte de terrorisme. L'utilisation des méthodes pour le couplage d'enregistrements pour empêcher le terrorisme peut avoir des impacts sur la collecte des autres données en termes d'augmentation de la non-réponse, et diminution de la qualité de la donnée. La plupart des efforts de collecte des données gouvernementaux incluent les garanties d'intimité, de la protection des renseignements personnels et de la non divulgation. Cette présentation est axée sur les relations entre les avancements dans le domaine du couplage d'enregistrements et le risque de divulgation et les efforts contre le terrorisme servant des grandes bases de données.

MOTS CLÉS : Fusion de données; contrôle de la divulgation; appariement de fichiers; contrôle d'influence; protection des renseignements personnels

1. INTRODUCTION

The purpose of this paper is to discuss possible roles for statisticians and statistics in counterterrorism (CT) efforts. Particular emphasis is placed on record linkage. The paper will review some advances in topics, such as record linkage theory, latent class modeling, and confidentiality research. A goal of the paper is to connect disparate topics such as CT and access to research data and by raising questions add to the scope of the discussion of these issues.

The author is neither an expert on global CT nor an expert on privacy and international law and does not possess a security clearance. Readings are limited to sources on the Internet and newspapers, magazines, and journals. The opinions expressed are those of the author alone and the frame of reference is predominantly based on the U.S.

Section 2 of the article first discusses statistics and CT and delineates some possibilities. Section 3 describes record linkage and the use of latent class models. Section 4 presents ideas concerning the use of and limits to record linkage in CT efforts. Section 5 contains comments on issues of privacy and confidentiality. Section 6 summarizes ideas on the relationship between CT and access to research data. Section 7 is a summary and presents some suggestions from the author.

2. STATISTICS AND COUNTERTERRORISM

Statistics concerns inference and prediction from data and the assessment and control of uncertainty. Counterterrorism (CT) involves efforts to discover, prevent, and intercept terrorist plots and activities. According to the U.S. National

[1] Michael D. Larsen, 220 Snedecor Hall, Iowa State University, Ames, Iowa 50011, USA, larsen@iastate.edu

Counter Terrorism Center (2006) Worldwide Incidents Tracking System, there were 11,000 terrorist attacks in 2005 with 14,500 casualties. Fifty-six casualties occurred in the U.S. Thirty percent occurred in Iraq and half involved no casualties. Unfortunately for monitoring, the definition of a terrorist attack was changed from the previous year, making trending difficult or impossible.

At least three factors likely matter in terms of the ability of statistical methods to aid CT efforts. One factor is the size of the threat: a large-scale 9/11-type plot, a medium- or small-sized group of individuals planning an action, or a lone individual. A second factor, related to the first but not exactly the same, is whether the activity is local, regional, national, or international in scope. A third factor is whether or not the area is a region of conflict. Techniques appropriate for one scope of prevention might not be helpful for others.

In the case of a very small scope (perhaps a lone individual) in a local area in a non conflict area, CT efforts could be helped by enhanced law enforcement, better databases on criminals and potential weapons (guns, components of explosives, etc.), better communication procedures and technology, and better security at government, other public, and critical sites. Statistics potentially can contribute to these efforts through better database matching software, search algorithms for scanning databases across jurisdictions, summaries of relevant suspect data, and threat prediction modeling. In all cases, uniformity in databases and compatibility among sources of data so that they can be used more readily can be advocated. Such improvements aid law enforcement in general (e.g., *Ames Tribune* 2006). Monitoring of international phone calls, data mining of airline databases, and tracking of international monetary transactions, however, are unlikely to be effective in such scenarios.

The rest of this section will focus on the scenario of an attack in the U.S. at the medium or large scale perhaps with international connections. CT statistical possibilities include (1) record linkage, (2) data mining, (3) social network analysis, (4) biological surveillance, (5) cyber attack detection, and (6) threat evaluation, prediction, and risk analysis. Examples of (1), (2), and (3) will be given below. In all cases, from a statistical perspective, there is a need for inference, not just description, and attention to common statistical problems.

Record linkage involves bringing together various databases on a population, matching individuals to records on two or more databases based on fields of information such as names, date of birth, and address, and assessing whether or not records across databases pertain to a single individual. If there were lists of buyers of trucks and farm chemicals and lists of individuals with contacts to various groups of concern due to international connections or hostile language in publications, then one could potentially match the lists and identify individuals common to both lists. This could motivate an investigation. Of course, one must assess whether the link is real or simply coincidental, whether the connection if it is real is truly suspicious, and how important is the potential threat.

Data mining uses statistical techniques to find patterns in a database that might be either common or unusual (see, e.g., Two Crows Corporation 2005). Unusual patterns in appropriate databases might correspond to suspicious activities. Databases on travel, phone conversations, types of jobs, purchases, etc., if they could be appropriately linked, might contain some patterns of activity that might be worth investigating. Besides limitations of databases and the ability to link them, data mining for CT requires inference concerning the revealed patterns in the information and insight into plans and potential scenarios to reduce the number of uninteresting associations.

Social network analysis looks for patterns in social networks, which are defined by phone, email, mail, travel, and financial connections. One use of such analysis is to find connections among groups of suspected individuals and identify others who could be in their group. Of course, inference is needed regarding the quality of the network information (who actually is contacting whom?), whether or not the group exists except as a coincidence of associations, and whether or not the supposed group poses any real threat. Tracking domestic and foreign telephone calls without listening to content, however, is not likely to uncover secretive terrorist plots (Bergstein 2006; Keefe 2006).

There are several statistical aspects that are common to the CT statistical applications. First, inference is required, especially because it is of interest to detect more than the 'sure' cases; one must weigh the evidence and decide what to do about it. Second, the degree of association or predictive ability could be quite weak given the available data: demographics including ethnicity and religion really will not be that predictive in general.

Third, there is the possibility of serious errors, errors can be costly, and measuring errors can be difficult or impossible. False positives occur when a person or group is falsely accused of terrorist activity. Numerous individuals abroad and in

the U.S. have been held falsely for long periods of time, treated horribly, and released without charges (Hegland 2006a, 2006b, 2006c; Cole 2006). False negatives occur when a terrorist activity, such as transfer of money, stealing of weapons, and violent acts are not detected or prevented. Depending on definitions there are numerous undetected acts.

Fourth, there are many potential barriers to implementing statistical CT methods. There could be legal barriers to data access, political barriers due to territoriality of agencies or political restrictions on activities, technological and staff limitations, issues of database incompatibility, and resistance by individuals against invasion of privacy.

Fifth, the quality and (un)availability of data significantly limit what can be accomplished. It is reasonable to assume that terrorists would want to avoid detection and take steps to do so. Apparently, it is not that hard to avoid detection in the U.S. Over 1% of the U.S. population is not counted in the decennial census. Sex offender registries are not up to date and locations of sex offenders are unknown. Parole violations are numerous. The number of illegal immigrants is in the millions. Yet many individuals in the above mentioned groups have government identifications, receive government benefits, or go to or send children to public schools. There is neither a central U.S. statistical agency/data repository nor is there default sharing of data across agencies or levels of government (not that there should be). Lack of centralization and privacy laws limit feasible, legal monitoring of the population. Further, even if many data sources were available to some part of the government (legally or not), the predictive power of the information in terms of accurately predicting an imminent terrorist threat is not necessarily high at all. Data can be out of date, recorded in error, and non unique.

In summary, the problem is hard. The CT investigators, politicians, and law enforcement officials need to be told how hard it is to implement these methods in the face of problems and severe risk of false positives. The methods are not fool proof. To reduce the chance of error and improve the likelihood for correct inference, the best statistical methods should be used. It will be very hard to test methods. Building realistic databases to test methods will be a challenge – few true positives are known – and methods should be "robust" to departures from prior assumptions.

Statisticians can advocate for improvement and voice their concerns. Three options for improvement are as follows. Redefine problems: instead of trying to search the world, try more focused searches (where there are better data), especially those for which warrants are defendable. Standardize government data: this will help if linkage is ever attempted and produce other possible benefits in terms of doing research. Improve data quality: this will help with linkages and searches as well as standard government functions and research.

Statisticians should voice their concerns about confidentiality pledges and access to research data. Confidentiality pledges must be upheld: data are gathered for specific purposes and people provide data under this belief – violating it will severely limit ability to gather data, and severely reduce privacy. See, for example, Fienberg (2004). Access to research data should not be substantially compromised: economics, education, health, medicine, science, and social science all need data. Perceived abuses for CT work could lead to a reaction that limits data access and prevents legitimate research.

The next few sections discuss methods of record linkage, record linkage and CT in more detail, and issues of nondisclosure in this regard.

3. RECORD LINKAGE AND LATENT CLASS MODELS

Record linkage (RL) is one method that can be used with counterterrorism (CT). RL also has other uses (Alvey and Jamerson 1997, Larsen and Rubin 2001, and references therein, Cormier 2005, Krewski et al 2005). In the basic application of RL, two files on a single population are compared in order to identify the individuals represented by records on both files. The situation is challenging when there are not unique identification (ID) numbers or codes, the files are large, and information is recorded with possible errors or is partially missing. When the files concern individuals, the matching variables besides SSN or other ID numbers can include first name, last name, middle name or initial, street name, house number, unit number, telephone number, relation to head of household, age, sex, race/ethnicity, and other variables available on and comparable across both files.

In common to many record linkage operations are the following factors.

1) Large scale; computerized information.
2) No unique, error-free, always-present ID number; SSN is a good start.
3) Use common variables in files to judge similarity: Age/date-of-birth, sex, race, address, phone, relation to head of household.

4) Errors in data; pre-processing of data is critical.
5) Process is repeated many times; prior information and insight are available.
6) Blocking of records, similar to stratification, is used to reduce computation and to increase accuracy.

In many record linkage operations, comparisons are reduced to sets of binary (1/0, yes/no) comparisons, one comparison for each field of information being compared to judge match status. A score is computed for each pair of records, one from each file. Points are added for agreement. Points are subtracted for disagreement. Rules are defined to handle missing data and partial agreement on character or number fields. Points are adjusted for the rarity (or commonness) of attributes. If the total score for a pair is high enough, then it is declared to be a matched pair. If the total is low enough, then the pair is declared to be a nonmatch. If resources permit, some records could be sent to clerical review for further data gathering or human intervention in the matching decision. Often, however, finding certain matches is the objective, and clerical review might only be used for those cases that are almost at the match declaration score, if at all.

Fellegi and Sunter (1969) proposed a statistical framework for consideration of record linkage. They provided mathematical justification for viewing the record linkage problem in terms of choosing cut-off scores for match and nonmatch in order to not exceed pre-specified error levels. False matches occur when pairs declared to be matches in fact refer to different people. False nonmatches occur when pairs that correspond to the same person are declared to be nonmatches. Fellegi and Sunter (1969) also suggested ways to estimate appropriate amounts to add or subtract based on agreement or disagreement, respectively, on the available fields of information.

Latent class models, and mixture models more generally, can be applied when the population under consideration can be viewed as arising from subpopulations. The set of pairs considered in a record linkage operation can be viewed as arising from a set of matches and a set of nonmatches. Latent class and mixture models can produce estimates of the needed scores for use in Fellegi and Sunter's (1969) algorithm. They also can directly estimate error rates and rank pairs of records based on their likelihood of being matches (Larsen and Rubin 2001 and references therein).

Significant statistical development has been accomplished in record linkage by viewing the problem as one of estimating components of latent class models. See Winkler (2006), Lahiri and Larsen (2005), Gomatam et al (2002), Winglee, Valliant, and Scheuren (2005), and Krewski et al (2005) for examples and sources.

Despite significant advances that have been made in RL theory and applications, RL operations can only be successful if powerful matching variables are available with few errors. Further, it is necessary to have some true cases produced either through clerical review or better linkage information (such as SSN) so that the performance of RL operations can be evaluated. Without training and test cases, one cannot be certain that automated computerized RL procedures actually are identifying matches and nonmatches.

In summary, many important issues can be studied through RL. RL, in some applications, can be facilitated by use of latent class models. Software and theory advances have made RL an even more powerful research tool than it was in past decades. Within agencies and in controlled settings, confidentiality of subjects can be protected. RL is hard work in major applications, requires significant careful development, and should be carefully evaluated to ensure high quality linkage results.

4. RECORD LINKAGE AND COUNTERTERRORISM

Record linkage (RL) for counterterrorism (CT) purposes could proceed in three steps. First, collect multiple databases or lists. Second, match them to find individuals common to two or more lists. Third, analyze the results by examining the resulting matches. In the first step, the multiple databases could consist of, for example, lists of suspected terrorists and their associates, travel manifests (airlines), criminal files (law enforcement), car/van rentals (commercial), weapons purchases (federal reporting), and purchases of bomb making material (commercial, federal reporting).

In the second step, matching could be accomplished using latent class models and various high capacity systems, such as those developed by the U.S. Census and Statistics Canada. Various scoring and matching alternatives would likely produce different results. For each pair of lists being compared, it would be necessary to define new match criteria based on available fields. The third step requires looking at the information from matches and potential matches and using human judgment and insight to consider the matching results and their implications for CT. See also Gomatam and Larsen (2004).

A variety of results would likely occur depending on the lists used and matching criteria. It is possible that one would find criminals. If enough matchable and relevant information were available, then one could potentially find suspected terrorists. Of course, one might identify people traveling to the Middle East or with relatives there. In any case, one would likely identify a lot of matches that mean nothing for CT. Further, there likely will be a lot of possible matches that actually do not match.

The output from a RL operation will be lists of pairs of records from the files that are similar in important ways – but are not necessarily the same person or indicative of terrorist planning. For example, suppose there is a Muslim farmer from the Middle East living in Iowa who purchases farm chemicals and trucks, or someone with a similar name who visits family in her/his home country once every other year and has unpaid parking tickets, or someone with a similar name who visits the same countries and has a lot of contacts in U.S. mosques and Muslim religious circles. Depending on matching criteria, some of the individuals described above could be identified.

Lenient matching criteria will create multitudes of potential matches. The practical difficulty of examining so many potential links is daunting. Strict matching criteria will lead to missed matches and amplify the impact of database errors. In short, a lot of human review needed and then a lot of investigation will be needed in order to try to use RL on a general population for CT purposes. As a result, it is likely that general RL of large databases will have low potential for CT. It will be challenging to improve matching algorithms without test data (data for which true status is known). It will be challenging to simulate realistic data for tests of procedures.

If RL is applied to large population lists without strong identifying information and very relevant predictor variables, a lot of false positives will result. The potential negative impacts of such a result are loss of privacy, lower participation in surveys, and distrust of government.

There are steps that could be taken to make general population RL more effective. First, use several blocking and matching criteria and pass through each list multiple times. By doing so, one field of information is not used too strictly. Second, adapt name information/scoring algorithms for the target population. Generally this has been investigated for Hispanic and some Asian populations in the U.S. Third, remove as many non-interesting potential matches as possible. Fourth, follow leads and do purposeful searches of suspected terrorists and criminals. A specific search likely will be more effective and more justifiable legally.

Unfortunately, it will be difficult for statistical inference procedures to help very much in a formal sense – there is not much potential to form a realistic test database on which to evaluate rare chance occurrences or estimate error rates. Despite that limitation, statistical reasoning still is important. The reminder to investigators that chance events do occur, misspellings happen, and nothing in the databases can prove by itself that someone is a terrorist or plans a terrorist action. A better understanding of RL, what it can and can not do and what affects its operation could be helpful to investigators who must decide what to do with RL results.

In general, the steps that would help RL for CT would help improve law enforcement and sex offender registries, organized crime and illegal drug prevention activities, Census enumeration, and epidemiological research. Standardization of how databases record information would increase the potential for linkage and searching. Study of RL options will require test files with known match status of considerable size and complexity. Also, education of practitioners of RL and large-scale data mergers about blocking strategies and scoring systems for RL could lead to overall improvements in statistical information systems.

5. NONDISCLOSURE, PRIVACY, AND CONFIDENTIALITY

Several terms are used in the statistical and government literature for protecting the identity of respondents to surveys and of individuals with information in administrative databases. These terms include nondisclosure, disclosure control, disclosure avoidance, privacy protection, confidentiality, and inference control. According to the *Statistics Canada Privacy Impact Assessment Policy* (Statistics Canada 2006):

> **"Privacy** is the right to be left alone, to be free from interference and from intrusions. It includes the right of individuals to determine when, how and to what extent their information is shared with others. The collection of information from respondents by Statistics Canada is, by its nature, a privacy-intrusive activity."

> **"Confidentiality** denotes an implied trust relationship between the person supplying information and the individual or organization collecting it. The relationship is built on an assurance that the information will not be disclosed without the person's permission. Under the *Statistics Act,* information that would identify individuals, businesses or institutions cannot be disclosed without their knowledge or consent."

Some regulations concerning privacy and confidentiality apply to entire nations. HIPPA (Health Insurance Portability and Protection Act) in the U.S. requires written consent to gather health and medical information; see, for example, http://www.cdc.gov/privacyrule. In general, however, policies in the U.S. are decentralized: each agency and state/local government has its own rules (or not). Despite the decentralization, the situation is that there are numerous pledges of privacy and confidentiality that apply to all versions of data collected for government surveys and stored in government databases.

Much of the data gathered by the government could also be of great use to researchers studying the U.S. population. Data gathered in government surveys, reports of cancer cases and other diseases, and socioeconomic data from the census all form the basis of important research. Thus, many studies would be possible if data were released to researchers for analysis. Sometimes researchers receive data through special restricted use licenses that typically involve establishing an adequate data security plan. In other cases, data are subjected to processing so that the version of the data that is released is deemed safe in terms of respondent confidentiality.

Many methods have been developed in recent years to protect respondent confidentiality. A first step is to remove names, addresses, and other key identification variables from the files before release. Several second steps that can be used alone or in combination include the following:

- Cell suppression methods in tables of counts so that very small counts in some cells do not identify individuals with unique characteristics;
- Top coding of variables, such as income, so that extreme values cannot be linked to specific individuals;
- Collapsing into cells/interval coding so that unique values of quantitative variables and small geographic locations do not correspond to single or a few individuals;
- Random noise, data perturbation methods that add random values to measurements so that unique values do not have a clear correspondence to individuals in the population;
- Data swapping methods that switch characteristics (female/male, minority/non minority, rural/urban, etc.) for pairs of respondents so that the linking power of matching variables is diminished;
- Sampling from the records instead of reporting on the whole population or sample so that uniqueness of records in the released data does not necessarily correspond to uniqueness in the population;
- Remote access server systems that report data summaries based on submitted queries without delivering the micro-level data to the researchers; and
- Synthetic data methods that generate artificial but realistic information based on the actual data, so that researchers have a complete set of micro-level records for analysis but no actual data are released.

Books such as those by Willenborg and de Waal (1996, 2001), Doyle et al (2001), Domingo-Ferrer (1998, 2002), Domingo-Ferrer and Torra (2004), and Domingo-Ferrer and Franconi (2006), special issues of the *Journal of Official Statistics* (1993, 1998), a report of the Federal Committee on Statistical Methodology (1994), and several technical reports at NISS (http://www.niss.org) have recently considered these issues and methods for making data available to researchers, but preserving confidentiality.

Developments in record linkage theory and methods have implications for confidentiality protection methods. Specifically, it can be possible in some situations to identify some individuals in publicly released databases after efforts have been made to make them anonymous. Winkler (2004a, 2004b) has considered techniques for measuring the risk of re-identification of individuals after efforts at de-identification have been implemented. More generally, Duncan and Lambert (1989), Skinner and Elliot (2002), and Reiter (2005) have examined way to measure the risk of disclosure in released databases. The tradeoff between the utility of the released data and the disclosure risk was the focus of Karr et al (2006).

Privacy and confidentiality usually are promised to individuals who provide data to the government and to research organizations. Failing to protect these individuals is a violation of the law and their trust. Advances in record linkage mean that care must be taken when releasing data and data summaries. Sophisticated techniques have been and are being developed to allow data to be available for research. Implications of counterterrorism activities based on record linkage and data mining for access to research data are discussed in the next section.

6. COUNTERTERRORISM AND ACCESS TO RESEARCH DATA

Both counterterrorism (CT) efforts and research require data. If CT efforts violate confidentiality or appear to violate promises regarding privacy, then it could make getting data harder. It seems reasonable to assume that people will increasingly refuse to give consent. Indeed, in response to perceived abuses, laws could be passed to increase protection and restrictions on use of data. A few examples of efforts to collect data for CT that have met with criticism are described below. The purpose of this article is not to examine the effectiveness or legality of various CT efforts.

The U.S. government wanted to collect information on what people read (e.g., bomb making manuals, information on national infrastructure). Some local libraries have refused to comply or destroyed records so that they cannot be taken. The effort generated a lot of negative publicity. The provision for collecting this information was rarely used and eventually removed from the renewal of U.S. Patriot Act.

It has been reported in the news that the U.S. government has been monitoring international phone calls. The 1978 U.S. Foreign Intelligence Surveillance Act prohibits this spying without special court approval (Bradley et al 2006). The government can even delay getting a warrant until 72 hours *after* wire tapping. Few of the millions of calls per year have been judged worthy of eavesdropping (Basu 2006). Further, the effort was concealed from Congress: initially it was denied, then it was said to be small scale, but it was widespread. A lot of negative publicity surrounded this effort.

Domestic phone calls also have been tracked. This would seem to violate laws on privacy that should protect U.S. citizens. The courts were not used and efforts were concealed from Congress (Broder 2006). This has lead journalists to speculate that these efforts, as much as being about CT, are about trying to establish unlimited presidential power to have wiretaps without warrants. That is, if President Bush and his administration argue that they have the right to do these activities and are not stopped, then will it become de facto policy in the U.S. (Hersh 2006, Drew 2006)? The purpose of this article is not to argue this point one way or the other.

Other monitoring efforts have been reported in the news. These include monitoring of electronic records, including emails and Internet activity, (Lichtblau and Risen 2005), reporters to finds sources of "leaks" (Drew 2006), and radiation levels at homes and mosques (*Chicago Tribune* 2005). A couple of monitoring efforts have been discontinued due to combinations of objections and lack of success. These include the collection of travel manifests and airline ticket information as part of Homeland Security's Computer-Assisted Passenger Pre-Screening program and the Pentagon/DARPA's Total Information Awareness (TIA) program. See also James (2006) and Miller (2006).

One additional method that the U.S. government has used to gather data is through *national security letters* (Schmitt 2005; Gellman 2005). They are used by the FBI to get information from telephone companies, libraries, Internet service providers, banks, credit bureaus, and other businesses. Historically about 300 were issued per year. Recently the number issued has been in the thousands per year. Data collected based on the letters apparently do not have to be destroyed after they are first evaluated and can be shared with others. It is not clear what privacy laws might be relevant once data are collected based on one of these letters.

Can general surveillance be accomplished without violating privacy? It seems unlikely given the concerns that have been raised concerning specific programs. In order for the government to use the data effectively for counterterrorism purposes, the data must have personal identifiers to be able to link it. The government must link data sets of highly personal information in order to find something useful. Further information on intent, social contacts, purchases, and movement are needed as well. This would seem only possible if the executive branch were granted unlimited power. See also Fienberg (2004).

In the name of counterterrorism, several actions have been taken by the U.S. government that could make people less willing to provide data to the government and to researchers. There is little or no evidence that there has been effective counterterrorism as a result of each of these actions. Many people have been arrested falsely or detained. Statisticians should oppose actions that threaten the ability to do research and use record linkage for legitimate purposes.

7. SUMMARY

Some statistical methods can be useful for counter terrorism, but *general screening* of the population (e.g., data mining, record linkage, and social network analysis) is very likely to be (1) ineffective, (2) costly, (3) productive of many false

positives, (4), a violation of privacy laws and confidentiality guarantees, and (5) a negative impact on research data quality and access.

Record linkage (RL) has many important uses. Latent class models can be used to estimate probabilities involved in some record linkage procedures. Serious large-scale RL uses require a lot of specialized work and attention to details of populations. Efforts could be made to improve the institutional RL capability. New advances in theory and software suggest increasing usefulness of RL. RL possibilities will be hindered if data quality suffers and access decreases.

Based on the considerations presented in this article, a few suggestions can be made to governments. First, get warrants and follow the law. Second, prefer targeted searches over wide scans. Third, improve databases and standardize information formats. Fourth, improve law enforcement and government legal information sharing (Marsh 2006).

A few suggestions can be made to statisticians as well. First, advocate for continued access to research data. Second, oppose government efforts that could lead to reductions in response rates and data quality. Third, raise questions about highly speculative statistical applications that produce large error rates with severe costs. Fourth, produce innovative and testable ways to use rich data sources in real time.

ACKNOWLEDGEMENTS

The author would like to thank the Survey Methods Section of the Statistical Society of Canada and its president, Patricia Whitridge, and the 2006 Program Chair, Richard Lockhart, for the invitation to speak on this topic at the 2006 meeting of the Statistical Society of Canada. The author also would like to thank the U.S. Census Bureau, the U.S. National Center for Health Statistics, and the National Institute of Statistical Sciences for support and interesting discussions. The opinions expressed in this article are those of the author alone and not necessarily of any other person or organization.

REFERENCES

Alvey, W., and Jamerson, B. (1997). *Record linkage techniques – 1997, Proceedings of an International Workshop and Exposition*, Washington, D.C.: Federal Committee on Statistical Methodology, Office of Management of the Budget.

Ames Tribune. (2006). DNA databank helps ID suspects. May 23, 2006, page B3.

Basu, R. (2006). So where's the justification for spying on Americans? *Des Moines Register,* February 8, 2006

Bergstein, B. (2006). Can mining data really uncover terrorists? *Des Moines Register*, March 25, 2006, pages 1D – 2D.

Bradley, C., Cole, D., Dellinger, W., Dworkin, R., Epstein, R., Heymann, P.B., Koh, H.H., Lederman, M. Nolan, B., Sessions, W.S., Stone, G., Sullivan, K., Tribe, L.H., and Alstyne, W.V. (2006). On NSA Spying: A letter to Congress. *New York Review of Books*, February 9, 2006, pages 42-44.

Broder, D. (2006). Congress must be consulted prior to use of secret wiretaps. *Des Moines Register*, February 9, 2006, page 11A.

Chicago Tribune. (2005). FBI watched homes, mosques for radiation. December 24, 2005, page 12.

Cole, D. (2006). Are we safer? *New York Review of Books*, March 9, 2006, pages 15-18.

Cormier, M. (2005). Record linkage overview. *Canadian Cancer Registry Manuals*. Health Statistics Division, Statistics Canada, Catalogue no. 82-224-XIE – No. 005. Ottawa, Ontario, Canada: Statistics Canada.

Domingo-Ferrer, J. 2002. Inference Control in Statistical Databases: From Theory to Practice (Lecture Notes in Computer Science). Springer.

Domingo-Ferrer, J., and Torra, V. 2004. Privacy in Statistical Databases: CASC Project International Workshop (Lecture Notes in Computer Science). Springer

Domingo-Ferrer, J., and Franconi, L. 2006. Privacy in Statistical Databases: CENEX-SDC Project International Conference (Lecture Notes in Computer Science). Springer

Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L. (2001). Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies. Amsterdam; New York: North-Holland/Elsevier.

Drew, E. (2006). Power grab. *The New York Review of Books*, volume 53, number 11, June 22, 2006, page

Duncan, G., and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business & Economic Statistics*, 7, 207-217

Federal Committee on Statistical Methodology. 1994. Report on Statistical Disclosure Limitation Methodology. Statistical Policy Working Paper 22. Office of Management and Budget, Washington, D.C.

Fellegi, I.P., and Sunter, A.B. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Fienberg, S.E. (2004). Homeland insecurity: Datamining, terrorism detection, and confidentiality. *NISS Technical Report*, 148. Research Triangle Park, NC: National Institute of Statistical Sciences.

Gellman, B. (2005). The FBI's secret scrutiny. *Washington Post*, November 6, 2005, page A01.

Gomatam, S., and Larsen, M.D. (2004). Record linkage and counterterrorism. Chance, 17 (1), 25-29.

Gomatam, S., Carter, R., Ariet, A., and Mitchell, G. (2002). An empirical comparison of record linkage procedures. *Statistics in Medicine*, 21, 1485-1496.

Hegland, C. (2006a). Empty evidence. *National Journal*, February 3, 2006.
http://nationaljournal.com/scripts/printpage.cgi?about/njweekly/stories/2006/0203nj4.htm

Hegland, C. (2006b). Guantanamo's Grip. *National Journal*, February 3, 2006.
http://nationaljournal.com/scripts/printpage.cgi?about/njweekly/stories/2006/0203nj1.htm

Hegland, C. (2006b). Who is at Guantanamo Bay. *National Journal*, February 3, 2006. http://nationaljournal.com/scripts/printpage.cgi?about/njweekly/stories/2006/0203nj2.htm

Hersh, S.M. (2006). Listening in. *The New Yorker*, Mary 29, 2006, pages 24-26.

James, F. (2006). Traveler monitoring triggers outrage. *Des Moines Register*, December 3, 2006, page 24A.

Journal of Official Statistics, 1993, Volume 9, number 2, Special issue.

Journal of Official Statistics, 1998. Volume 14, number 4, Special issue.

Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., and Sanil, A.P. 2006. A framework for evaluating the utility of data altered to protect confidentiality. AMER STAT 60 (3): 224-232.

Keefe, P.R. (2006). Can network theory thwart terrorists? *New York Times Magazine*, March 12, 2006, pages 16-18.

Krewski, D., Dewanji, A., Wang, Y., Bartlett, S., Zielinski, J. M. and Mallick, R. (2005). The effect of record linkage errors on risk estimates in cohort mortality studies. *Survey Methodology*, 31, 13-21

Lahiri, P., and Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 222-230.

Larsen, M.D., and Rubin, D.B. (2001). Iterative automated record linkage suing mixture models. *Journal of the American Statistical Association*, 96, 453, 32-41.

Lichtblau, E., and Risen, J. (2005), Nation's phones tapped. *Chicago Tribune*, December 24, 2005, pages 1 and 12.

Marsh, B. (2006). U.S. security: failures, near-failures, and an A-minus. *New York Times*, February 26, 2006, Op-Ed page 3.

Miller, L. (2006). Innocent travelers linked to terror lists. *Des Moines Register*, October 8, 2006, page 8A.

National Institute of Statistical Science. Technical reports (http://www.niss.org)

Reiter, J.P. 2005. Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100 (472): 1103-1112.

Schmitt, R.B. (2005). Patriot Act debate was misguided, some fear. *Des Moines Register*, December 12, 2005, page 4A.

Skinner, C.J., and Elliot, M.J. 2002. A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B*, 64: 855-867, Part 4.

Statistics Canada. (2006). Statistics Canada Privacy Impact Assessment Policy. http://www.statcan.ca/english/about/pia/piapolicy.htm

Two Crows Corporation. (2005). *Introduction to Data Mining and Knowledge Discovery, Third Edition.* Potomac, Maryland: Two Crows Corporation. http://www.twocrows.com/intro-dm.pdf

U.S. National Counter Terrorism Center. (2006). *Reports on Incidents of Terrorism 2005*. April 11, 2006. http://wits.nctc.gov/reports/crot2005nctcannexfinal.pdf.

Willenborg, LCRJ, and de Waal, T. 1996. Statistical Disclosure Control in Practice. (Lecture Notes in Computer Science). New York: Springer

Willenborg, LCRJ, and de Waal, T., and Willenborg, L. 2001. Elements of Statistical Disclosure Control. (Lecture Notes in Computer Science). New York: Springer

Winglee, M., Valliant, R., and Scheuren, F. (2005). A case study in record linkage. *Survey Methodology*, 31, 1, 3-11.

Winkler, W.E. (2004a). Re-identification methods for masked microdata. *Privacy in Statistical Databases, Annals of the New York Academy of Sciences*, *Proceedings,* 3050: 216-230.

Winkler, W.E. (2004b). Masking and re-identification methods for public-use microdata: Overview and research problems. *Privacy in Statistical Databases, Annals of the New York Academy of Sciences*, *Proceedings*, 3050: 231-246.

Winkler, W.E. (2006). Overview of record linkage and current research directions.
U.S. Census Bureau, Statistical Research Division, Research Report 06-02. Washington, D.C: U.S. Census Bureau.