Journal of Privacy and Confidentiality (2011) **3**, Number 2, 57–63

Commentary: Future U.S. National Statistics Use of Administrative Data

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*

George T. Duncan*

Government statisticians are intrigued by the possibility of accessing administrative data as a way to enhance survey and census data. Survey data are rich in attributes, but they exhibit sampling errors. More consequentially, they demonstrate non-sampling errors, often because of nonresponse, but also because of responses that are incomplete or inaccurate, frequently due to the respondent's inability to recall. Additionally, panel data are subject to attrition. Administrative data can help compensate for these problems importantly because administrative data, such as social security files, include almost everybody. Further, supplementing statistical data can be cost-effective, because an enormous amount of administrative data are already collected to support the functional operations of government agencies, and so are potentially available. In many instances, respondent burden in surveys can be lessened because certain attribute values may be available from administrative records. Some U.S. federal agencies, e.g., the Social Security Administration (McNabb et al., 2009) have linked administrative data with survey data to broaden its demographic and socioeconomic measures and also to improve the quality of the survey data. See National Research Council (2005), pp.45–48 for additional examples and discussion of benefits of such data linkage.

Increased data sharing of personal data among federal agencies for statistical purposes has been recommended provided confidentiality can be protected; see National Research Council (1993), pp. 98–99. Drawing on his long experience with the U.S. Census Bureau, Gates (2011) establishes both what is known about confidentiality risks in this context and what needs to be known so that the potential uses of administrative data for statistical purposes can be realized. The issues he concentrates on are far-ranging, and they are ones that must be addressed in policy making.

For my part, I pose two questions. The first question concerns future changes in the meaning of administrative data. The second question concerns public understandings of privacy and confidentiality and especially the issue of consent by the data subject. For each question I will briefly explore some implications of the answers for policy of federal statistical agencies.

*Professor Emeritus, H. John Heinz III School of Public Policy and Management and Department of Statistics, Carnegie Mellon University, Pittsburgh, PA mailto:gtduncan@gmail.com

© 2011 by the authors

1 Is the Scope of "Administrative Data" Enlarging?

Gates (2011) identifies administrative records as profiles of interactions of individuals and businesses with government. This definition should be regarded as broader than might at first be thought, certainly because private-sector interactions are increasingly subject to government reporting requirements, beginning for an individual with birth and ending with death (and even not quite then, given estate tax filing requirements), and beginning for a business with license applications and, regrettably for some, ending with bankruptcy filings. Further, local and state government agencies provide administrative data to the federal government. The Social Security Administration (SSA), for example, has matched benefit and earning reports with files identifying homeless people compiled by New York City's Department of Homeless Services. SSA used the linked data to produce statistics showing the impact of benefits and earnings on the homeless population's use of shelters. As another example, members of the National Cancer Registry provide lists of cancer patients to SSA, and industry epidemiologists provide SSA with industry-specific lists of former employees. These files are used to check SSA's death records, beneficiary rolls, and earnings files to ascertain if the persons have died or can be presumed alive. SSA also links its administrative data with survey data from the National Health Interview Study. See McNabb et al. (2009).

The scope of administrative data is expanding in large part because of employment of data types that are not traditional to surveys and censuses. These data types include geospatial, audio and video, biometric recognition, biological material, and electronic network. Not because it is the only relevant new data type to our topic, but simply for reasons of focus, I will confine my discussion to two types: geospatial data, which integrates maps with numerical attributes, and electronic network data, which captures interactions in social media sites, such as those among Facebook friends. I will note some confidentiality issues specific to each.

Duncan et al. (2011), p. 155 state, "Importantly for social research, the linkage of spatial data with personal data has much value in developing understanding and public policy in social, economic, political, environmental and public health realms." Dissemination of geospatial data poses particular problems of disclosure risk because of detailed geographical specificity. For example, Brownstein et al. (2006) demonstrated how the minimum resolution of publication of figures in the New England Journal of Medicine could permit a high percentage of addresses to be identified. National Research Council (2006) discusses the progress that has been made in developing restricted data approaches to lowering disclosure risk as well as restricted access and synthetic data approaches. In the specific area of public health, Golden et al. (2005) explore confidentiality issues involving geospatial data.

Social network data once was the domain of researchers studying interactions in relatively small groups or organizations, and these data were laboriously collected based on phone calls, personal visits, and the like. Today, e-traffic at social media sites yields billions of interactions among millions of nodes, all automatically captured electronically. Kleinberg (2007) cites examples of how researchers have used such data to study topics such as community formation, collective decision making, and opinion influences. What

are potential uses of such data by policy makers? Two applications that come to my mind are in the employment statistics area and in the consumer price statistics area.

Today, job seekers are increasingly employing search strategies through social media sites such as LinkedIn (other online job sources include Monster, CareerBuilder, Craigslist, SimplyHired, Dice, and Vault, but they do not have the widespread social media aspects that LinkedIn does). Agencies might incorporate such data into current protocols to give more sophisticated assessments of the size of the pool of job seekers, and hence obtain a better understanding of the unemployment rate. Of course, appropriate adjustments must be made for differences in the demographics of such data and that of the general population. For example, LinkedIn visitors are a substantial oversample of Asians and an undersample of Hispanics, and also an oversample of those with a graduate school education and an undersample of those with no college. From the other side of the employment equation, employers are increasingly making known the availability of job openings through social media. Data on changes in the level and composition of such openings can provide evidence that would be useful to economic planners.

In the consumer price area, the Bureau of Labor Statistics contracts with the U.S. Census Bureau to carry out the Quarterly Interview Survey and the Diary Survey. Data on transaction prices from those online sites that bring together buyers and sellers might provide cost-effective ways of getting information about realized prices on a wide range of products. For example, eBay, with its millions of auctions running simultaneously, has a humongous database of potentially relevant prices. eBay Marketplace Research provides, as a subscription service, a complete databank of eBay buying and selling statistics, including quantities, dates, and individual prices.

These examples clearly suggest a broadened scope for the administrative data that would be useful for statistical purposes and public policy making. Such data may involve hundreds of millions of data subjects from whom data are automatically obtained through everyday transactions. What then about privacy and confidentially issues? In particular, what about the basic principle of consent?

2 Can a Data Subject Realistically Consent?

What decision making role in the CSID data process (Capture, Storage, Integration and Dissemination; see Duncan (2004)) should a data subject play? What are the ethical and practical factors that should structure the role of the data subject in consenting to the use of the data they provide? Good answers require an understanding of the relationship of privacy and confidentiality, specifically the implications of data subject perceptions about them.

Privacy pertains to individuals' ability to control information about themselves (Boruch and Cecil, 1979). As such, privacy is a personal construct intimately coupled with individual autonomy. In contrast to privacy, confidentiality refers to how personally identifiable information is managed and disseminated. Simply put, privacy

concerns people and confidentiality concerns data. In our context, confidentiality is a policy issue for federal statistical agencies, but one that depends critically on how confidentiality policies are perceived by data subjects. The distinction most relevant here is that confidentiality requires protection of personal data rather than control of data collection.

But this distinction between privacy and confidentiality is becoming less germane. Propelled by computer technology, the acts of data collection and data dissemination are becoming harder to separate. Data gathered today can in more and more cases be disseminated essentially instantaneously. To a large extent, therefore, information policy makers will need to simultaneously consider privacy and confidentiality as inherently intertwined. See Duncan et al. (2011), pp.149-150).

For a data subject, the act of consenting to the use of their data can be either active or passive. Active consent, whether written or verbal, requires a positive response. Passive consent occurs when data subjects are notified about the intent to use their data. They are told that no objection on their part will be taken as consent. An objection, however, can prevent the use of the data. This is different than notification where data subjects are just informed that the data they provide will be used for specified purposes.

Since consent suggests that the individual has a viable alternative—most starkly to opt out entirely, this aspect of privacy is complicated when data subjects may be either legally compelled to provide information (e.g., tax filing) or have substantial incentives to do so (e.g., welfare or loan applications). For the data subject, therefore, an opt-out option is not a realistic possibility. For this reason, conditioning program participation on the completion of blanket information release consent forms is not voluntary and hence not real consent. See Preis (1999) for such an argument in the context of a parent seeking mental health treatment for their child.

But is it realistic for an administrative agency to provide an opt-out option? If the data to be collected are directly germane to the administrative function, this issue can be viewed as of no substantial ethical consequence and so need not be addressed. However, when administrative data are to be shuffled on to a statistical purpose, the issue does indeed need to be addressed. National Research Council (1993), pp. 4–5 note that privacy and confidentiality considerations commend the concept of *functional separation* between statistical and administrative data. In 1977 the Privacy Protection Study Commission affirmed that data that have been collected by government agencies exclusively for statistical and research purposes should not be moved to administrative purposes. But what are the concerns about movement in the opposite direction, that is, from administrative purposes to statistical purposes?

Consider Fair Information Practices (see, e.g., Gellman (2011)), one principle of which is that information should only be used for the purpose for which it is collected. Would statistical uses of administrative records be in violation of the principle? Perhaps because Fair Information Practices are endorsed less in the United States than in Europe, they could be considered irrelevant to the situation addressed by Gates. But, since this principle does have rather general appeal on ethical grounds, how should it be addressed? One stance is that the principle precludes statistical uses unless the

data subject is informed of potential uses for statistical purposes and has the power to preclude such uses. Implementing such a stance is burdensome on both the agency and the data subject, but has been done in certain specific instances. For example, the University of Michigan's Health and Retirement Study, because of limitations under the Privacy Act, can link to SSA administrative data only if the survey respondent has signed a release.

Importantly, requiring consent could potentially increase the cost of data collection and lower the utility of the data so much that it would cripple statistical uses. Also, to use prior administrative data would require data subjects to be re-contacted, which would not be uniformly possible and presumably could be prohibitively expensive. A more realistic approach, that still provides attention to the ethical dimensions, would require the approval of an independent body for transfers of data from administrative purposes to statistical purposes.

Fundamental to discussions of the ethics of consent is that it be informed. In information collection situations, this stricture means that an individual should possess relevant knowledge about what information will be shared, with whom, how it will be used, and for how long. If administrative data are to be routinely used for statistical purposes, this new use will need to be communicated effectively to the data subject. Certainly this communication cannot be realistically assumed based on the data subject reading the Privacy Act requirement of a System of Records Notice or searching the websites of agencies for their privacy statements. But meaningful communication is challenging because in many situation it is essentially impossible to describe all uses to which the data will be put, and blanket phrases such as "statistical purposes" are too vague. These arguments suggest that further research is needed on strategies for agencies to effectively communicate to data subjects about statistical data use of the administrative data they provide and what their options are.

References

Boruch, R. and Cecil, J. (1979). *Assuring the Confidentiality of Social Research Data*. University of Pennsylvania Press.

Brady, H., S., G., Powell, M., and Schink, W. (2002). Access and confidentiality issues with administrative data. In *Studies of Welfare Populations: Data Collection and Research Issues*. National Research Council. `http://aspe.hhs.gov/hsp/welf-res-data-issues02/08/08.htm\#administrative\%`.

Brownstein, J., Casa, C., and Mandl, K. (2006). No place to hide—reverse identification of patients from published maps. *New England Journal of Medicine*, 355(16): 1741–1742.

Duncan, G. (2004). Exploring tension between privacy and the social benefits of governmental data bases. In Podesta, J., Shane, P., and Leone, R. (eds.), *Little Knowledge: Privacy, Security, and Public Information after September 11*, 71–88. New York: The Century Foundation.

— (2007). Privacy by design. *Science*, 317: 1178–1179.

Duncan, G., Elliot, M., and Salazar-González, J. (2011). *Statistical Confidentiality: Principles and Practice*. Washington, DC: Springer.

Gates, G. (2011). How uncertainty about privacy and confidentiality is hampering efforts to more effectively use administrative records in producing U.S. national statistics. *Journal of Privacy and Confidentiality*, 3(2):3–40.

Gellman, R. (2011). Fair Information Practices: A Basic History. `http://bobgellman.com/rg-docs/rgFIPshistory.pdf`.

Golden, M., Downs, R., and Davis-Packard, K. (2005). Confidentiality Issues and Policies Related to the Utilization and Dissemination of Geospatial Data for Public Health Applications. Technical report, The Socioeconomic Data and Applications Center (SEDAC) Center for International Earth Science Information Network (CIESIM) Columbia University, New York.

Kleinberg, J. (2007). Challenges in mining social network data: Processes, privacy, and paradoxes. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07*. New York.

McNabb, J., Timmons, D., Song, J., and Puckett, C. (2009). Uses of administrative data at the social security administration. Social Security Bulletin, Vol. 69 No. 1.

National Research Council (1993). Private lives and public policies: Confidentiality and accessibility of government statistics. In Duncan, G., Jabine, T., and de Wolf, V. (eds.), Panel on Confidentiality and Data Access, Committee on National Statistics. Washington, DC: National Academies Press.

— (2005). Expanding access to research data: Reconciling risks and opportunities. In Panel on Data Access for Research Purposes, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academies Press.

— (2006). Putting people on the map: Protecting confidentiality with linked social-spatial data. In Guttman, M. and Stern, P. (eds.), Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data. Committee on the Human Dimensions of Global Change, National Academies Press.

Preis, J. (1999). Confidentiality: A Manual for the Exchange of Information in a California Integrated Children's Services Program. Sacramento: California Institute for Mental Health.