Research power and confidentiality issues associated with combining data sets

Geraldine P. Mineau
Department of Oncological Sciences and
Huntsman Cancer Institute
University of Utah

Jean E. Wylie
Utah Resource for Genetic and Epidemiologic Research
University of Utah

Ken R. Smith
Department of Human Development and Family Studies and
Huntsman Cancer Institute
University of Utah

Contact Information:
Geraldine P. Mineau
geri.mineau@hci.utah.edu
2000 Circle of Hope
Huntsman Cancer Institute
Salt Lake City, UT 84112

Paper prepared for: Population Association of America, March 29-31, 2007
New York City, NY

Abstract: Considerable benefits are associated with electronically linked datasets for demographic, biomedical and health researchers. Although such databases have substantial benefits, the public's concern about misuse and the need for appropriate safeguards have sparked national discussions. This paper addresses some of the issues that are associated with combining or linking data sets for demographic research and how they can be addressed. It focuses on the challenges and solutions regarding the relationship with the data contributor. The development of research resources or centers to take the role of linking data sets for research use is explored. The Utah Population Database (UPDB) will be used as an example of one possible way to approach these issues. It has existed for over 30 years and has a long history of developing policies and procedures that address such issues. It provides access to about 9 million records and supports 65 projects.

Extended Abstract

Background

Considerable benefits are associated with electronically linked datasets for demographic, biomedical and health researchers (Wylie and Mineau, 2003). Continued population-based research is essential and especially useful in these disciplines. Although such databases have substantial benefits, the public's concern about misuse and the need for appropriate safeguards have sparked national discussions. This manuscript will address some of the issues that are associated with combining or linking data sets for demographic research and how they can be addressed. There are many examples of the value of this activity: cohort studies that need follow-up information, reconstructing families and kin-networks in historical demography, health studies that link patient records with outcomes, to name but a few. The Utah Population Database (UPDB) will be used as an example of one possible way to approach these issues.

The UPDB is a rich source of information for genetic, epidemiological, demographic, and public health studies. For over 30 years, researchers have used this resource to identify and study families that have higher than normal incidents of cancer or other diseases, to analyze patterns of genetic inheritance and to identify specific genetic mutations. In addition, demographic studies have shown trends in the fertility transition, changes in mortality patterns for both infants and adults, and identified characteristics of exceptionally long-lived individuals. The UPDB provides access to about 9 million records and supports about 65 research projects. The central component of UPDB is an extensive set of Utah family histories, in which family members are linked to medical information.

There are five data contributors for the records housed in the UPDB; these records are linked and merged to create an infrastructure that may be used by researchers. The contributors

are the Utah Department of Health which provides all vital records (birth, marriage, divorce, death and fetal deaths), the Family History Library of the Church of Jesus Christ of Latter-day Saints, Utah Cancer Registry, Cancer Data Registry of Idaho, and Utah Driver License Division of the Utah Department of Public Safety. In addition there is a link to the University of Utah Health Sciences enterprise data warehouse which houses all medical information for hospitals and clinics. Lastly, we have worked with a specific research project that has data from the Centers for Medicare and Medicaid Services (CMS) including both the vital status file as well as person-specific files with medical information.

Privacy and Confidentiality

One of the main concerns regarding use of linked (or merged) datasets is the issue of protection of the identity of individuals named in these data. Privacy is the ability to control information about oneself while confidentiality is the obligation of a second party to not reveal private information about an individual to a third party without the permission of the person concerned (Wylie and Mineau 2003). Given the public's need for confidence in research activities ". . . policies protecting privacy and confidentiality in research . . . are essential not only to protect individuals but to endure the advancement of science. (Phimister 2001)."

Removing explicit identifiers, such as name, address, and Social Security number, has been used to insure confidentiality before releasing information to researchers on the assumption that the resulting data look anonymous. Even when certain of these steps have been taken, Sweeney (1997) and Malin and Sweeney (2004) point out that there are methods of re-identification, such as matching to other data bases or by looking at unique characteristics found in the fields of the database itself resulting in so-called deductive disclosure. Furthermore, for

matching individuals across data sets, removing identifying information is not desirable thus other approaches need to be employed.

Relationships with Data Contributors: Challenges and Solutions:

In this paper, we will address three common issues in working with data contributors.

- Under what conditions may data, not collected specifically for research, be re-used for research without compromising the privacy of the individuals named in the data? Typically individuals are not contacted and asked consent regarding the use of their records, thus the privacy concept has been waived or put aside by the data contributor. Does the provider of the data have the authority to allow research use of individual-level data including linking to other data? What guarantee or safe guards are given to the provider regarding the confidentiality of the individuals named in data.
- Are there limitations or restrictions regarding research use? Do the data providers have the right to review use and publications? Is there an appeal process that can be used by the researcher?
- Who owns the data? Do the data remain (under agreement) the property of the provider while the value-added component of the linked data sets and derived information belong to the research effort?

Separation of research resource from research use

Certain kinds of research infrastructures or statistical coordinating centers, associated with Universities or (not for profit) public institutes, may take the role of linking data sets for

research use. Such centers or resources need to develop policies and procedures associated with the release of information. Some of the more fundamental procedures in use now include:

- Policies that prevent researchers from linking to other data resources (without approval) to prevent the inference of individual information outside the scope of the original research agreement. (Kohan end Altman, 2005)
- Procedures to create de-identified data sets after matching and before releasing individual-level information for research use. This includes the development of web-based queries for data-mining that follow HIPAA rules
- Protocol (or not) for contacting individuals for the purposes of recruitment into a research protocol. It may be important to have a protocol for contact and recruitment of individuals into a research study which can collect biosamples, request the release of medical records and collect self-reported information. "Privacy and confidentiality can be achieved by policies that require third-party contact of, and consent by, individuals before researchers receive identifying information on individuals. (Wylie and Mineau, 2003)"

Other examples of similar resources will be discussed, for example from the field of genomics, the concept of the Charitable Trust has been suggested (Winickoff and Winickoff, 2003).

Other topics that may be emphasized are the increased importance of a relational database and whether the research projects return information to the resource. While these are related more to procedure than policy, they address the need to 1) retain the source of information associated with each record or field, 2) keep sensitive data in separate tables with restricted access, and 3) keep source documents separate from derived data or returned data. This type of assurance ties into the relationship with the provider of data.

Conclusions: There has to be a balance between creating more powerful research resources with the costs associated with them and the need to develop policies and procedures to ensure ethical research practices.

References

Kohane, I. and Altman, R. (2005) Health-Information altruists—a potentially critical resource. *New England Journal of Medicine* 353: 2074-2077.

Phimister, B. (2001) Protecting individuals and promoting science. *Nature Genetics* 28: 195-196.

Sweeney, L. (1997) Weaving technology and policy together to maintain confidentiality. *J Law Med Ethics* 25: 98-110.

Malin, B. and Sweeney, L. (2004) How (not) to protect genomic privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *J Biomed Inform* 37:179-192.

Winickoff, D. and Winickoff R. (2003) The charitable trust as a model for genomic biobanks. *New England of Medicine* 349:1180-1184.

Wylie, J. and Mineau, G. (2003) Biomedical databases: protecting privacy and promoting research. *Trends in Biotechnology* 21: 113-116.

Acknowledgements

We thank Huntsman Cancer Foundation for database support provided to the Utah Population Database. Websites are maintained for these resources at http://www.utah.edu/rge and http://www.hci.utah.edu/groups/ppr