Understanding Identifiability in Secondary Health Data

Niko Yiannakoulas, PhD

ABSTRACT

Secondary health data are becoming important sources of information for health research. Unfortunately, there are few national guidelines outlining clear criteria for data custodians when adjudicating requests for access to these data. This presents a particular challenge to researchers and even public health officials interested in multi-jurisdictional analysis of health data. While some of the problem in developing consistent national standards may stem from differences in provincial privacy law, it may also be the result of imprecise terminology, particularly around the concept of identifiability. This commentary discusses the differences between identifiability and self-identification with the hope of contributing to a broader discussion establishing national guidelines for the use of secondary health data in research.

Key words: Data protection; data sharing; medical informatics

P rovincial administrative health records can be important sources of secondary health data for researchers. Not only do these data yield larger sample sizes, but they may also give information about health diagnoses that differ from self-report surveys (with neither being a true “gold standard”). Secondary data allow researchers to compare health across subgroups of the population, study patterns or risk factors over time and understand influences on disease that interact to produce complex effects that cannot be detected with other data. When combined with recent advances in computer technology and statistical methodology, secondary health data are a rich resource for potential research questions in a wide variety of fields. While some secondary data are released in a form that is publicly accessible, most researchers must request access to these data through data custodians in addition to going through a process of ethics review. In Canada, these secondary health data custodians typically include hospitals and provincial health ministries governed by provincial and federal privacy legislation. This legislation is often specific to health data,1 and in the case of provincial legislation, usually covers multiple intra-provincial jurisdictions.2 While there are national guiding principles for research ethics boards to determine ethical use of secondary data,3,4 there are few national guidelines for data custodians to follow when assessing the suitability of releasing data for research. Part of this may be due to differences in provincial privacy legislation,5 however, the existence of national terms of reference for the use of secondary health data6 suggests a widespread acknowledgement of a need for national standards.

Recently, we made an application to a data custodian outside our province of residence for access to small area cancer data. Our interest was in analyzing municipality-level incidence of several types of cancer for persons 65 years of age and older. In our request, the only attributes associated with each observation were year and municipality of residence. While the data we requested were temporally and geographically aggregated, small populations made it possible that some municipalities would have fewer than 5 cancer cases in some years. The data custodian refused our request on the grounds that they did not release aggregate data with such small counts. Their primary stated concern was the identifiability of individual stakeholders.

According to the Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans (TCPS), the use of secondary data with identifying information requires ethical scrutiny ensuring, among other things, that the identifying data are necessary for research and that the privacy of participants will be protected.1 Unfortunately, there has been insufficient clarity or formal discussion in Canada about when the disclosure of secondary health data meets the criterion of identifiability, or more importantly, whether or not identifiability is always a reasonable cause for privacy and confidentiality concern. This is a serious challenge to conducting national-level research using secondary health data, which already confronts the problem of navigating different provincial privacy laws2 and variations in the application of ethical guidelines.7

Identifiability versus self-identification

Identifiable data are data that contain detail sufficient to identify a single individual, exclusive from any other individual, with a high degree of certainty. Unique or direct identifiers (such as social insurance numbers) are highly identifiable, since the probability that two people have the same number is very low (and in fact only possible through error or fraud). A quasi-identifier is a combination of

Author Affiliations

Correspondence: Niko Yiannakoulas, Assistant Professor, Health Geomatics Lab, School of Geography and Earth Sciences, McMaster University, Hamilton, ON L8S 4K1, Tel: 905-525-9140, ext.20117, Fax: 905-546-0463, E-mail: yiannan@mcmaster.ca

Conflict of Interest: None to declare.
attributes in a data set (such as age, sex and postal code) that are not independently identifiable, but together make persons identifiable either directly, or through linkage to another source of data.\(^8\) Importantly, while identifiability and privacy are related concepts, identifiable data are a privacy concern only when they are linked or linkable to other information that could be considered private, sensitive or that contribute to the identification of an individual. A list of social insurance numbers by themselves does not represent a disclosure of private information; indeed, such a list can be generated randomly. Were the list to also include corresponding names and birth dates, public disclosure of such data could be used to cause serious harm. This illustrates an important distinction between identifiability and self-identification; some data contain sufficient information for individuals to identify themselves while at the same time remaining anonymous to the general population. A person who identifies their own social insurance number from a list of social insurance numbers may self-identify, but remains unidentifiable to anyone else looking at the list. Similarly, the public disclosure of a single case of a rare illness in a small community does not publicly disclose the identity of the case, but the individual diagnosed, his or her family, and the health care providers involved would likely know to whom the case referred.

The release of small area health data most often involves disclosing self-identifiable rather than identifiable data more generally. Consider the following amendment to our original data request. Say that for each municipality with case counts less than 5, the real count was replaced with a random number between 0 and 5. On occasion, the random number generated could equal ‘1’ in a town that has, by chance, one case of disease. When this occurs, the one individual in the town with the disease could self-identify with this randomly generated case the same way he or she would if the case were derived from real data. However in both instances, the individual is merely self-identifiable, and no private information is being publicly disclosed.

The distinction between identifiability and self-identification is equally important in other applications. Consider the following list of information: male, 45 years of age, diagnosed with AIDS in 1997 and having a residential postal code starting with “T6E”. It is quite likely that this person is uniquely self-identifiable from these data attributes; that is, no other person possesses all of these attributes combined, so some individual could see this record and self-identify. However, as in the case above, this person is not identifiable in the population – his identity would almost certainly not be disclosed even if such a digital record entered the public domain. As is typical for self-identifiable data, the key information that would identify the individual (i.e., diagnosed with AIDS in 1997) is also private and known only to him and others to whom he discloses such information.

The confusion between identifiability and self-identification may stem from a failure to distinguish between the private health information and the identifiers/quasi-identifiers contained within secondary health data. Private health information is, by definition, not an identifier or quasi-identifier. Private information is information that is not generally known, and therefore, the release of such data, even as small counts in small geographic areas, does not represent a disclosure of identifiable information, no matter the sensitivity of the health information. Releasing such data may still be of ethical concern; it is possible that disclosing a single case of a disease in a small town could have undesirable social repercussions – such as motivating gossip, watchfulness, or other socially intrusive behaviour. Such social risks are very likely a function of population size, and of less concern in a community of 100,000 than in a community with a population of 100. While important issues, these social risks should not be confused with a disclosure of private information.

It is our perception that a lack of clearer national guidelines defining identifiability is an important obstacle to multi-jurisdictional research with secondary health data in Canada. The precautionary principle seems to govern the decisions of many data custodians, and while a convenient decision-making paradigm in the face of existing privacy legislation, it lacks the rigorous pretexts for properly assessing the costs and benefits of using secondary health data.\(^9,10\) While considerable research has been done to develop methods for usefully de-identifying health data,\(^11-13\) it is still important to have consistent and well-articulated guidelines for allowing and prohibiting the release of secondary health data for research purposes. This not only ensures that the risks of unwanted disclosure are minimized, but also ensures a transparent, informed and nationally consistent system for adjudicating requests to use these data.

Further national dialogue is required to establish clearer and more predictable standards for releasing secondary health data for research, and in particular, defining identifiability in terms that are truly relevant to privacy protection. Equivalent efforts are underway in the United States, requiring cooperation among a mixture of private and public stakeholders.\(^14\) Considerable national discussion on data privacy issues has occurred within the context of research ethics,\(^15\) however, data custodians and researchers must continue to participate in this process in order to ensure the development of clear and practical terms of reference. This could do much to advance national-level research, particularly since other data sources (such as the long-form Census) may become less reliable in the future.

REFERENCES

7. Willison DJ. Data protection and the promotion of health research: If the laws are not the problem, then what is? Health Care Policy 2007;2(3):39-43.

Dear Editor:

I am writing on behalf of the Global Health Research Initiative (GHRI) Steering Committee in reference to Larson et al.’s important commentary on Grand Challenges Canada.1 In the commentary, the authors refer to GHRI, a partnership of five federal government agencies and departments that has been funding global health research since 2001. Larson et al. underscore the importance of GHRI and its capacity to address important global challenges in health through balanced funding across a wide research spectrum, from biomedical and clinical research to research on health systems and services and the socio-cultural, environmental factors that affect the health of populations. They also mention that GHRI “funding is now mostly spent” and indicate that “its renewal is still in doubt”.

The GHRI Steering Committee would like to clarify that the partner agencies and staff are hard at work drafting a five-year strategic plan with the aim to renew the partnership’s Memorandum of Understanding (MOU) later this year. Preparatory work has involved an extensive consultation process with GHRI-supported researchers; other stakeholders in Canada and in low- and middle-income countries (LMICs); and representatives from the GHRI partners and other Canadian government agencies and departments.

The new GHRI strategic plan will build on the successes of this 10-year Canadian federal partnership. GHRI-funded work contributes to worldwide efforts to improve disease surveillance, immunization efforts, and prevention and response to health threats of regional and global significance, including emerging infectious diseases such as pandemic influenza, non-communicable chronic diseases, and vaccine-preventable diseases. In the last three years, GHRI launched three new programs that are now funding over 20 LMIC-Canadian teams. Since its inception, GHRI has nurtured equitable South-North collaborations by supporting nearly 200 Canadian researchers collaborating with over 500 LMIC researchers and decision-makers in over 60 countries, working on a wide range of global health research projects while building research capacity and promoting knowledge translation.

The GHRI partners are dedicated to supporting research for improved global health through partnerships between Canadian and LMIC researchers and decision-makers, and fostering important learning across boundaries, be they North-South, South-South or South-North.

Diane Gagnon, PhD
Program Leader
Global Health Research Initiative
Ottawa, Canada

On behalf of the GHRI Steering Committee

REFERENCE


Editor’s Note: This letter was sent to the authors of the commentary, who declined the invitation to respond.