Canadian Journal of Public Health (2020) 111:761–765
https://doi.org/10.17269/s41997-020-00303-8

INNOVATIONS IN POLICY AND PRACTICE

Less than five is less than ideal: replacing the "less than 5 cell size" rule with a risk-based data disclosure protocol in a public health setting

Krista Wilkinson[1] · Christopher Green[1] · Deborah Nowicki[1] · Christina Von Schindler[1]

Received: 23 April 2019 / Accepted: 11 February 2020 / Published online: 11 March 2020

Abstract

Setting The Winnipeg Regional Health Authority (WRHA) is one of the largest and most diverse health regions in Canada. Within the WRHA, the Population and Public Health (PPH) Surveillance Team provides epidemiological support across a variety of public health service areas.

Intervention We developed and deployed a risk-based data disclosure protocol that balances the need to share public health surveillance data with the need to protect personal health information.

Outcomes Unlike the conventional data disclosure standard adopted in Manitoba (suppress cell sizes < 5), the new protocol is based upon a risk-based re-identification approach that focuses on the size of the denominator instead of the numerator. This approach has allowed for innovation in data dissemination infrastructure within the unit that would not have been possible previously, including the deployment of public-facing cloud-based interactive maps and dashboards. It has also resulted in strengthened protection of personal health information as the risk of re-identification can now be precisely calculated across all data release situations.

Implications In challenging the "cell size less than five" rule, this project is an example of how a scientifically based data disclosure protocol can support a public health organization in meaningful sharing of population health data with community partners and the public. This helps ensure that program and policy responses are empirically based, strategically focused, and cross-jurisdictionally coordinated.

Résumé

Contexte L'Office régional de la santé de Winnipeg (ORSW) est l'une des régions de la santé les plus vastes et diverses du Canada. Au sein de l'ORSW, l'équipe de la surveillance de la Santé de la population et du public (SPP) fournit du soutien épidémiologique dans divers secteurs des services de santé publique.

Intervention Nous avons élaboré et mis en œuvre un protocole de divulgation des données fondé sur les risques qui tient compte du besoin de partager des données de surveillance de la santé publique et du besoin de protéger les renseignements personnels sur la santé.

Résultats À la différence de la norme conventionnelle de divulgation des données adoptée au Manitoba (élimination des cellules de valeur < 5), le nouveau protocole est fondé sur une approche de réidentification basée sur le risque qui met l'accent sur la taille du dénominateur plutôt que du numérateur. Cette approche a permis d'innover l'infrastructure de diffusion des données au sein de l'unité, laquelle n'aurait pas été possible auparavant, y compris la mise en place de cartes et de tableaux de bord interactifs publics axés sur l'informatique en nuage. Cette approche a également fourni une protection accrue des renseignements personnels sur la

Electronic supplementary material The online version of this article (https://doi.org/10.17269/s41997-020-00303-8) contains supplementary material, which is available to authorized users.

✉ Krista Wilkinson
kwilkinson2@wrha.mb.ca

[1] Winnipeg Regional Health Authority, Winnipeg, Canada

santé puisque le risque de réidentification peut dorénavant être calculé avec précision dans toutes les situations de divulgation des données.

Répercussions En remettant en question la règle de « cellules de valeur < 5 », ce projet représente comment un protocole de divulgation des données fondé scientifiquement peut appuyer un organisme de santé publique dans le partage significatif de données de santé de la population avec des partenaires communautaires et le public. Les réponses en matière de programmes et de politiques sont ainsi empiriques, stratégiques et coordonnées de façon interjuridictionnelle.

Keywords Privacy · Data disclosure · Personal health information · Public health surveillance · K-anonymity · Cell suppression · Re-identification risk

Mots-clés Confidentialité · Divulgation des données · Renseignements personnels sur la santé · Surveillance de la santé publique · K-anonymat · Élimination des cellules · Risque de réidentification

Introduction

To effectively address the many complex and "wicked" (Rittel and Webber 1973) public health issues that cross jurisdictions (homelessness, problematic substance use, emerging infectious diseases, chronic disease epidemics, etc.), it is critical that our public sector organizations find ways of breaking out of their data silos to meaningfully share information with relevant stakeholders, including community partners and the public. The integration of data from diverse jurisdictions can enable the development of a comprehensive picture of an issue in terms of its root causes, a description of populations at greatest risk, and an exploration of the most effective and impactful interventions. A comprehensive knowledge base can help ensure that program and policy responses are empirically based (not based on gut feeling or partial information) and are as strategically focused, cross-jurisdictionally coordinated, and cost-effective/impactful as possible.

One of the barriers to sharing data at a meaningful level of granularity has been adherence to privacy legislation protecting personal health information (Fairchild et al. 2007). Cell size suppression rules are used widely across Canada; these approaches require that cells containing non-zero counts of less than a specified (arbitrary) number be suppressed. A recent example of the data suppression landscape in Canada is seen in a national report on apparent opioid-related deaths; the suppression preferences of provinces and territories varied in not only count size but also specific attributes (Special Advisory Committee on the Epidemic of Opioid Overdoses 2019). Although easy to communicate and operationalize, evidence supporting suppression rules is lacking. Suppression rules do not guarantee that disclosure of personal health information will not occur (Matthews et al. 2016) and can prevent the release and sharing of meaningful data even when denominators are large and the risk of re-identification is very low.

When the emphasis is focused primarily on the protection of personal health information, less attention is paid to the opportunity costs of not making data widely available for supporting critical program and policy decisions. Concerns around privacy have led governments to adopt uncompromising data policies, at the expense of the improvements that could come as a result of sharing data sources (Macek 2019). The benefits of protecting the privacy of an individual need to be balanced against possible risks to public health.

To address these issues, the Population and Public Health (PPH) Surveillance Unit in the Winnipeg Regional Health Authority (WRHA) undertook the development of a modernized and flexible data disclosure protocol in a public health setting. This paper describes the process of how PPH Surveillance staff worked with their organizational privacy officer to implement a formalized risk re-identification approach to privacy and data release.

Setting

The WRHA provides healthcare for the more than 770,000 residents of Manitoba's largest urban centre. Within the WRHA, surveillance and epidemiology services are delivered as a component of PPH. Routine and ad hoc surveillance data are collected across multiple areas, including communicable diseases, healthy parenting and early childhood development, and healthy sexuality and harm reduction.

In Manitoba, *The Personal Health Information Act* (PHIA) is the privacy law that establishes the rules governing the collection, use, and disclosure of personal health information. The WRHA is considered a trustee under *The Act* and has an obligation to protect the privacy of individuals for whom it collects information (Government of Manitoba 2017). The general surveillance practice in the WRHA for sharing data publicly has been to provide aggregate tables in static reports, with cell sizes of five or fewer individuals suppressed.

Intervention

Literature review and consultative process We conducted an iterative literature review and consultative process between October 2016 and December 2018 to inform protocol development. We did not do a formal systematic review as we expected our understanding and focus to evolve throughout the discovery period. We initially searched both the academic and grey literature for articles in the general themes of privacy, de-identification, and anonymization. The reference lists of relevant articles were further searched by hand to identify other papers of potential interest.

We did extensive consultations within the health region, including recurring sessions with the WRHA Chief Privacy Officer, the Director of WRHA Ethics Services, and research associates from the Centre for Healthcare Innovation at the University of Manitoba. We also consulted stakeholders in Health Information Managements and Analytics as well as Epidemiology and Surveillance within the Manitoba provincial government. We met with researchers at the Manitoba Centre for Health Policy (University of Manitoba) and did further consultations with industry experts and with individuals from various public health agencies across Canada.

We used the findings from the literature and consultative review to develop the data disclosure protocol and associated tools. The protocol and tools underwent review and revision multiple times as our understanding of the field matured.

Outcomes

The key documents we used when developing the WRHA PPH Data Disclosure Protocol are listed in the Appendix in the ESM. During the process of literature review and consultation, risk-based re-identification emerged as a defensible and transparent method of protecting privacy and our protocol was ultimately built on a framework with the size of the denominator as the basis for assessing re-identification risk (El Emam 2010). The choice of this framework was substantively informed by a report from the Ontario Information and Privacy Commissioner (Information and Privacy Commissioner 2016) as well as from colleagues at the BC Centre for Disease Control who had already implemented a similar approach. We were unable to identify any literature providing scientific rationale that supported the "cell size less than five" method.

We adopted a risk re-identification approach based on the *K*-anonymization statistic (1/*K*) which assumes that the probability of being able to identify an individual row (representing an individual person) in a published dataset is indirectly proportional to the number of other rows in the dataset sharing the same quasi-identifier (El Emam and Dankar 2008). A quasi-identifier is a piece of information that by itself does not identify an individual, but can allow re-identification when combined with other pieces of information. For example, although birth date, gender, and postal code are not directly identifiable by themselves, they become potentially identifiable when combined (Sweeney 2000). All rows in the dataset sharing the same quasi-identifier values form an "equivalence class," with the probability of re-identification of an individual record equal to 1 divided by the size of its equivalence class.

For example, imagine a simple dataset that contains three quasi-identifiers (age—30 or 35; gender—male or female; geography—X or Y) where each unique combination of quasi-identifiers forms an equivalence class (Table 1). In this example, all 35-year-old men living in neighbourhood X form equivalence class V, which contains 2 members (subjects 9 and 10). The probability of using quasi-identifier information to correctly re-identify a record in that equivalence class is 1/2 or 0.5. As the size of the equivalence class grows, the probability of being able to correctly re-identify a record decreases; by the time the class size has grown to 20 members, the likelihood of re-identification has decreased to 5%.

Using our protocol, a surveillance dataset is first stripped of all direct/personal identifiers (name, PHIN, social insurance number, etc.), retaining only a set of quasi-identifiers (age group, gender, geography) which are used to calculate the equivalence class size as described above. If the minimum equivalence class size is 20 or greater, the dataset is considered to be a candidate for public release; if the minimum equivalence class size is <20, then an alternative anonymization strategy is selected depending on the specific utility requirements of the release. For example, a quasi-identifier may be dropped from the dataset (e.g., no longer stratified by gender

Table 1 Example dataset with three quasi-identifiers forming eight equivalence classes

Subject	Age	Sex	Geography	Equivalence class
1	30	Male	X	I
2	30	Male	X	I
3	30	Male	Y	II
4	30	Male	Y	II
5	30	Female	X	III
6	30	Female	X	III
7	30	Female	Y	IV
8	30	Female	Y	IV
9	35	Male	X	V
10	35	Male	X	V
11	35	Male	Y	VI
12	35	Male	Y	VI
13	35	Female	X	VII
14	35	Female	X	VII
15	35	Female	Y	VIII
16	35	Female	Y	VIII

or age group) or expanded in size (e.g., using a broader age group category) in order to achieve an equivalence class of 20 or greater. The dataset is iteratively modified in this manner until the re-identification risk threshold of 5% is achieved. The release threshold that we have adopted (1/20 or 5% chance of re-identification) is the industry-accepted threshold for the public release of data when the negative consequences of successful re-identification are considered to be extremely high and could result in significant potential harms and injuries to the individual (Information and Privacy Commissioner 2016). We also assume the likelihood of a re-identification attack on our data releases will always be 100%.

It is important to note that de-identification can never produce an aggregate dataset for which there is a zero probability of re-identification. Rather, as described, de-identification results in an aggregate dataset where the probability of deducing the identity of an individual is acceptably low given the benefits of releasing the data to end-users.

Along with completing the full re-identification risk assessment prior to releasing a dataset, we also included a review of the possible risks associated with attribute disclosure, defined as the disclosure of attributes relating to groups of individuals that may result in stigmatization and social harms. Developed with WRHA Ethics Services, the ethics of a particular data release are reviewed by public health content experts and management to ensure that any social harms associated with attribute disclosure are minimized. If potential harms are identified, the data release may be restructured to minimize these harms, or advanced communications and consultations may be undertaken with the affected populations to contextualize the data release.

We developed a suite of data disclosure tools to use alongside the protocol, including (i) a risk assessment template that quantifies the overall risk of re-identification, with documentation standards outlining when a formal assessment is required and where the resulting assessment is stored (Excel 2010 macro-enabled document), (ii) a brief guide to ethical principles and a checklist to inform attribute disclosure conversations, and (iii) generic code to use when assessing equivalence classes and identifying denominators of less than 20 individuals (Stata 13.0, College Station, TX).

The data disclosure protocol was endorsed by WRHA PPH management in December 2018. Following this endorsement, we deployed our first web-based interactive dashboard in early 2019 (Population Health Surveillance Team 2019).

Our protocol is meant to be applied only in situations where potentially identifiable personal health information will be shared publicly or broadly in aggregate form; it is not meant to apply in situations where disclosure is already permitted under PHIA. It applies to all public health reports, tables, summaries, or data extracts released publicly by the PPH Surveillance Unit. It also applies to aggregate data extracts uploaded into publicly accessible cloud-based tools.

Discussion

This paper describes how we developed and deployed a risk-based data disclosure protocol in the WRHA. Unlike the conventional cell size suppression rule, the new protocol is based on a risk of re-identification approach that focuses on the size of the denominator instead of the numerator.

Our data disclosure protocol has resulted in the strengthened protection of personal health information as the risk of re-identification can now be precisely calculated across all data release situations. Previously, suppression rules were violated when reporting rare events (such as a single case of measles in Winnipeg) where disclosure was considered a necessary part of the public health response. Our data disclosure protocol allows for the defensible sharing of rare diseases or events that would have a cell size of less than five at the regional level even without stratification by age and sex.

This approach has also allowed for innovation in the development of data dissemination infrastructure that would not have been possible previously, including the deployment of public-facing cloud-based interactive maps and dashboards. We needed a data disclosure protocol that would support the utility of the source tables for online dashboards. A particular problem of suppression rules is that of complementary suppression (e.g., the need to suppress other cells in the table as well as row and column totals to prevent back calculation of suppressed cell values). Complementary suppression can result in a number of "holes" in the data and lessen utility.

We believe that the most important benefit of this approach is that it facilitates the sharing of data between organizations at a meaningful level of granularity. The ability to systematically de-identify a dataset to meet an accepted re-identification release threshold, while at the same time maximizing data utility, enables the preparation of datasets that can be confidently shared (without unwarranted fear of identity disclosure) with other institutions and program areas for the purposes of planning and coordination, with academic partners for the purposes of research, and uploaded into low-cost cloud-based decision support tools.

We learned that it was crucial to engage early and often with our local privacy officer. The Chief Privacy Officer for the Winnipeg Regional Health Authority was initially very cautious when we first proposed moving away from the standard suppression method towards adoption of a risk re-identification framework. She challenged us to undertake more intensive literature searches and consultations in order to convince her of the merits of the new approach. This not only resulted in a more robust understanding of the re-identification landscape but also allowed our chief privacy officer to become an early champion of the protocol.

We also learned through the discovery process that data disclosure protocols need to align with organizational business needs and the privacy legislation to which they are

subject. For example, all personal information used by Statistics Canada is protected by both the Government of Canada's Privacy Act and the Statistics Act (Statistics Canada 2019); for Census data, Statistics Canada uses both area and data suppression, as well as random rounding to protect respondents' personal information (Statistics Canada 2016). In our local public health setting, using a similar rounding protocol to Statistics Canada would not allow for the reporting of small fluctuations in case counts that have meaningful implications for public health (for example, rounding down a single case of a rare disease to zero).

In Manitoba, PHIA clearly outlines when disclosure of information is permitted and mandates the protection of personal health information; however, it does not dictate how data trustees ensure that the required confidentiality is maintained. Two issues not addressed under PHIA but of concern to our stakeholders emerged through the consultative process: self-identification and attribute disclosure.

Our protocol does not apply to situations in which an individual re-identifies themselves or in which an individual may be recognized as forming part of a dataset by someone who already has that information about them. For example, it does not apply to situations where an adversary does not learn anything new about another person—if an adversary already knows all the information about another person included in a dataset (age, gender, geography, and disease information), then they would not learn anything new from the data release. In both these situations, no harms are created to individuals since the information was already known.

The other issue that generated a substantial amount of discussion was attribute disclosure. Although existing surveillance practice was to seek approvals from senior management and to engage with affected populations prior to release of potentially sensitive information, the data disclosure protocol formalizes the process and is systematically applied to all data releases. As our surveillance practice evolves, our intention is to engage with relevant stakeholder populations earlier in the surveillance cycle.

The focus of this paper was on how we pragmatically developed and deployed the risk re-identification approach to data disclosure in an applied public health setting. There are large programs of research and entire industries devoted to the protection of privacy and it was beyond the scope of this paper to go into more detail. We encourage readers who wish to gain a more nuanced understanding of the privacy field to consult the list of articles in the Supplementary material.

Conclusion

A scientifically based data disclosure protocol that uses established release standards to objectively balance the risk of re-identification against the benefits of data disclosure is critical to breaking down traditional data silos in government and for promoting an "open data" environment that enables the deployment of modern cloud-based decision support tools such as web-based interactive maps and dashboards. It equips public health organizations with the tools to share detailed and meaningful information legitimately and confidently with stakeholders to address emerging issues and helps avoid the significant opportunity costs of not using this information when making critical program and policy decisions.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

El Emam, K. (2010). Risk-based de-identification of health data. *Security & Privacy, IEEE, 8*(3), 64–67.

El Emam, K., & Dankar, F. K. (2008). Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association, 15*(5), 627–637.

Fairchild, A., Bayer, R., & Colgrove, J. (2007). Privacy and public health surveillance: the enduring tension. *The Virtual Mentor, 9*(12), 838.

Government of Manitoba. (2017). *The Personal Health Information Act*.

Information and Privacy Commissioner. (2016). *De-identification guidelines for structured data*. Ottawa.

Macek, C., & Boillot, N. (2019). *Opinion: The hidden costs of data protection in public health*.

Matthews, G., Harel, O., & Aseltine, R. (2016). Privacy protection and aggregate health data: a review of tabular cell suppression methods (not) employed in public health data systems. *Health Services & Outcomes Research Methodology, 16*(4), 258–270.

Population Health Surveillance Team. The epidemiology of communicable diseases in the Winnipeg Health Region, 2013–2018. 2019; Available from: https://public.tableau.com/profile/survdeploy#!/vizhome/CD-TSR2019_Reportreplica_FINAL/SUMMARYTABLE?publish=yes.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4*(2), 155–169.

Special Advisory Committee on the Epidemic of Opioid Overdoses. (2019). *National report: apparent opioid-related deaths in Canada (January 2016 to March 2019)*. Ottawa: Public Health Agency of Canada.

Statistics Canada. (2016). Census profile - area and data suppression. Available from: https://www12.statcan.gc.ca/census-recensement/2011/dp-pd/prof/help-aide/N3.cfm.

Statistics Canada. (2019). Privacy notice. Available from: https://www.statcan.gc.ca/eng/reference/privacy.

Sweeney, L. (2000). *Simple demographics often identify people uniquely*. Carnegie Mellon University, Data Privacy Working Paper 3: Pittsburgh.