"Real-world" de-identification of high-dimensional transactional health datasets.

Kenneth A. Moselle[a,1], Stan Robertson[b] and Andriy Koval[c]
[a,b]*Island Health;* [c]*University of Victoria*

Abstract. This paper presents a framework for addressing data access challenges associated with secondary use of high-dimensional transactional datasets that have been extracted from electronic health/medical records (EHRs). These datasets are subject to the data de-identification "curse of dimensionality" [1] which manifests as substantial challenges to preserving analytical integrity of data contents when high-dimensional datasets must be de-identified and deemed free of Personal Information (PI) prior to disclosure. A large array of methods can achieve this objective– for low dimensional datasets. However, these methods have not been scaled up to the types of high-dimensional data that must be sourced from the transactional EHR if the objective is specifically to generate products that can inform point-of-care clinical decision-making. The Applied Clinical Research Unit (ACRU) in Island Health is implementing a process that addresses key privacy challenges inherent in disclosures of high-dimensional transactional health data. This paper presents a schematic and abbreviated rendering of key principles and processes on which the ACRU approach is based.

Keywords. Privacy, data de-identification, curse of dimensionality, simulated data

1. Background - why the need for high-dimensional transactional health data?

Data required to generate clinically actionable products - evidence-based treatments are built on a foundation of information relating risk factors and interventions to outcomes for identifiable cohorts. For studies working from EHR data extracts, evaluation of treatment outcomes for acute conditions, and analytics supporting management of chronic and commonly co-emergent diseases will call for datasets containing a diverse array of cross-continuum clinical data contents that reflect treatment response or disease progression [2]. In the EHR, these contents (attributes of patients/clients; interventions) are layered onto a transactional data substrate depicting potentially large numbers of patient encounters with a broad array of programs that constitute a health service system, e.g., 1700+ programs in Island Health [3]. Given the multifactorial nature of clinical determinants and interventions, and the variably-spaced distribution of outcome-relevant information across an array of service encounters, efforts to generate clinically-useful and clinically targeted products from EHR extracts will translate routinely into requirements for large volumes of sparse high-dimensional data. Narayanan & Shmatikov [4] suggest this type of request is the rule, not the exception when working with "real world" transactional datasets.

Privacy challenges associated with high-dimensional datasets - even when direct identifiers (e.g., Name) have been removed from the EHR extracts, the highly granular, multivariate depictions of patient "journeys" embodied in the transactional service

[1] Director, Applied Clinical Research Unit, Island Health, contact kenneth.moselle@viha.ca

encounter data pose substantial challenges to privacy. The reason is that every patient trajectory within a sparse high-dimensional space is likely to be ***distinguishable*** at some level from all others. This translates into a privacy issue if the variables that distinguish cases in the dataset are available to the public in an identifiable form. When that is true, distinguishable cases in the dataset can be re-identified. Because distinguishability increases multiplicatively with the addition of new linked variables to a dataset, the privacy challenges associated with disclosure becomes very problematic in contexts that require datasets to be deemed free of Personal Information (PI).

Clusters of variables within a dataset that do not function as direct identifiers but nevertheless enable cases to be distinguished and re-identified are known as "indirect identifiers" or "quasi-identifiers – QIDs [5]. Risk for re-identification can be mitigated by rendering people "the same" with regard to distinguishing QIDs. If QIDs are rendered non-distinguishing by coarsening values (e.g., grouping diagnoses into superordinate categories) then this may not compromise analytical integrity of the data. But when re-identifiability issues are addressed by altering values on distinguishing variables in such a way that essential clinical "truth" is altered (e.g., if diagnoses are altered randomly for a subset of cases), then analytical integrity of the data is affected. Legacy de-identification methodologies have not provided a solution for high-dimensional transactional data that do not entail such perturbative alterations of data contents [6], which fundamentally compromise the utility of the data for point-of-service applications.

Legacy approaches to data de-identification - the challenge for a data de-identification methodology is to reduce risk for re-identification of nominally de-identified data while preserving analytical integrity of the data. Most methods that work from quantitative estimates of re-identification risk build from measures of distinguishability, and trace their roots back to a demonstration re-identification attack carried out by Sweeney [7]. Working with a small set of variables (full date of birth, full zipcode and gender) she showed that 87% of the records in the public registries she accessed were uniquely distinguishable. Her work provided substantial guidance for data de-identification methods sanctioned by the US Privacy Rule [8]. If those approved methods are applied to datasets composed of any of 18 classes of QIDs (and no other QIDs), then the risk for re-identification falls into the range .01% to .25% [9].

Any de-identification methodology that entails quantitative estimates of risk for re-identification of distinguishable cases will invoke some variation of Sweeney's approach at some stages in the de-identification process. However, various optimizations of her method (see Aldeen [6] for a thorough review) have not yielded an approach that preserves enough of the essential truth status of data, and derivative statistical properties, to warrant application of analytical results to the care of real patients.

The Applied Clinical Research Unit (ACRU) within Island Health, in partnership with various research groups at the University of Victoria, has taken on a program of research that is "forcing the issue" around developing a more complete set of tools that will enable researcher access to relatively pristine high-dimensional datasets, while meeting privacy obligations. A scenario-based method has been employed to generate a set of principles and procedures which provide a framework for the ACRU's contextualized approach to data access and data de-identification. What follows is a representative scenario, and an excerpt from the set of principles and procedures that are keyed to the privacy protection requirements associated with the data required to carry out the full suite of current ACRU research projects. The model is intended to support access on the part of the researchers located outside of Canada, so of necessity it targets the objective of rendering the data statistically and logistically/contextually free of PI.

2. A "real world" use case in the area of Mental Health & Substance Use (MHSU)

Clinical problem/research questions – what longitudinal trajectories are associated with excess morbidity and mortality in the MHSU population. What are the causes of this morbidity/mortality–all causes, including but not restricted to overdoses?

Information Products: rates of various outcomes and distal/proximal determinants; severity-stratified trajectories reflecting patterns of patient/client interaction with secondary and tertiary services (provided by the Health Authority); change-points in trajectories associated with service access and/or patient characteristics.

Data required: age, gender; transactions consisting of encounters plus dates with 1700 programs; acute care diagnoses, procedures (14,000 ICD9 categories); Emergency Department (ED) presenting complaints (165 values); ED Clinical Discharge Diagnoses; Minimum Reporting Requirements (Ministry of Health) for MHSU 346 variables, ≥ 1record per MHSU program registration, Pharmacy data; Vital Stats (deaths).

3. Data de-identification principles for "real world" data disclosure and use.

Principle #1 – work from a model of the data disclosure and use context. Stated in slightly different terms: create a *target information architecture* that identifies data sources, describes data movement, catalogues analytical approaches and intended products, and locates envisioned data product users/uses. This architecture points to processes where "source of truth" status is invested in the data. Specification of intended data users/uses provides a basis for setting utility constraints [10] on parameters around permitted/proscribed alterations to the data. In effect, the architecture provides an analytical "conscience" for de-identification activities that entail alterations of the data.

Principle #2 – Privacy Model: Distinguishability≠ Re-identifiability ≠ Risk for Re-Identification. "Privacy risk" can be unpacked into three entities: a) distinguishability, based purely on configurations of scores on QIDs within the dataset; (b) re-identifiability, which reflects a mapping of distinguishing pieces of data onto publicly available bodies of information; and (c) *risk* for re-identification. This last entity is a judgment that may be regarded conceptually as a multiplicative function of re-identifiability and a third component – a consideration of what data disclosure activities/ scenarios can be regarded as "reasonable" or "reasonably likely" on the part of the data user [11,12]. In attributing "risk" to a candidate data disclosure, it is essential to distinguish these three entities and to be clear about which of these three are being referenced.

Principle #3 – Risk-based access adjudication decisions or policies should be anchored in plausible scenarios. If publicized re-identification attacks are going to be invoked as a justification for data access policies or procedures, or if they are going to be discounted, then involved parties can presumably detail the relevance/comparability of those scenarios to the requested data disclosure. As well, they should be able to delineate a chain of reasonably likely activities that would result in re-identification. Barth Jones [13] and Cavoukian & El Emam [14] provide models for such analyses.

Principle #4 – Cost-benefit analysis. Quantitative estimates of risk based on distinguishability is not the only method for quantifying risk. Wan et al. [15] propose a more fully context-aware approach that employs game-theory and cost-benefit analysis, where the key question is whether there are any reasonably envisioned scenarios in which benefits outweigh risks. If there are no such scenarios, then there is zero risk – from a

cost-benefit perspective. Data access adjudicators will need to consider whether such an analysis will carry weight over more 'classic' estimates of risk based on distinguishability of cases. These 'classic' methods may appear to be more objective than cost-benefit approaches, but they introduce quantities (e.g., "adversary power") reflecting assumptions about the external context and data user knowledge that may not be regarded as correct or reasonable by knowledgeable parties [16].

4. De-identifying processes for high dimensional transactional data.

Process#1 – Examine the data at a univariate (single variable) level. Focus on distribution of values on variables in order to locate statistical outliers, where statistical risk may be "concentrated". As well, the data should be viewed qualitatively through data classification scheme "lenses" that are attuned to considerations such as differential sensitivity of data contents.

Process #2 – Data ecology – scan the environment. The objective is to identify publicly available datasets that would enable contents in the requested datasets to function as QIDs. This will substantially reduce the dimensionality of the dataset from a purely quantitative privacy risk perspective.

Process #3 – Secure the data disclosure environment and implement audit controls. Technical controls to prevent unauthorized access or import/export of raw person-level data protects against both current and difficult-to-envision future risks. Depending on what activities are tracked, the audit trail may shed light on at least a portion of the range of re-identification-relevant activities performed by data users.

Process #5a –If Processes #1-3 yield a small set of QIDs, and no issues around analytical integrity have been flagged: de-identify the data, *then* hand-off to researchers for analysis.

Process #5b – If approaches suitable for low-dimensional datasets put the analytical integrity of the data at risk - carry out the program of research analytics, *then* de-identify – or execute as an integrated process. The researchers are the parties that fit the data to statistical models and craft the data products. They are the parties that discover whether and where there are significant and useful contents in the data. As such, when analytical products are crafted from high dimensional data, it is reasonable to expect that workable utility constraints on changes to the dataset can only be determined ***after*** the researcher has worked with the data.

Process #6 – Generate simulated datasets, recruiting data de-identifiers, researchers, data scientists and domain experts. Such a team would collectively hold the knowledge and skill required to specify the minimum set of semantic and statistical properties that should be preserved in a simulated dataset, e.g., distributions at a univariate and multivariate level; treatment/exposure characteristics and designs to support real-world clinical trials; survival rates; clustering of variables; and a potentially broad array of other covariance relationships among variables [16].

Process #7 – Open/public release – and looping back to the original source data. Process #6, delivers a simulated version of a high-dimensional dataset that preserves ***some*** essential statistical properties and carries no privacy risk. Statistical models can be evaluated against these simulated data, and promising models can then be evaluated (under suitably controlled conditions) against the real data. What results will be validation/refinement of the models, as well as refinement of the simulated datasets.

5. Discussion

This paper proposes a methodology that entails a thorough examination and evaluation of the context of an envisioned data disclosure in service of ecologically informed decisions around privacy protection of the data. The methodology recruits the researcher as an early and "equal" partner in the process of setting utility constraints on data-altering activities employed to bring risk for re-identification down to a level deemed acceptable by data access adjudicators. The full model anticipates the use of emerging tools, e.g., Bayesian model discovery tools applied to multivariate databases [17] to generate simulated datasets that preserve more of the essential covariance and other properties of complexly structured data entities – and to support the open data agenda. These processes are not simple, but they are feasible. They hold out promise for breaking the curse of dimensionality – and enabling more thorough extraction of clinically useful content from the very large body of clinically pertinent information accumulating in real time in every location where an electronic medical/health record has been implemented.

References

[1] C. Aggarwal. On k-Anonymity and the Curse of Dimensionality. Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005 pp. 901-909.

[2] R. Sukkar, E. Katz, Y. Zhang, D. Raunig, and B. T. Wyman. Disease progression modeling using hidden Markov models. In EMBC, pages 2845-2848. IEEE, 2012.

[3] K. Moselle & A.Koval. Beyond administrative data and into clinical/transactional data: Introduction to the Cross-Continuum Coding Scheme (CCCS). Unpublished manuscript, January 25, 2017.

[4] A. Narayanan & V. Shmatikov, V. Robust de-anonymization of large datasets (How to Break Anonymity of the Netflix Prize Dataset) arXiv:cs/0610105v2 [cs.CR] 22 Nov 2007, pp. 1-24.

[5] K. Emam & L. Arbuckle. Anonymizing Health Data. O'Reilly, 2013.

[6] Y. Aldeen, M. Salleh & A. Razzaque A. A comprehensive review on privacy preserving data mining. SpringerPlus, December 2015, 694, pp. 1-36.

[7] L. Sweeney, L. Guaranteeing Anonymity when Sharing Medical Data, the Datafly System. Masys, D., Ed. Proceedings, American Med. Informatics Assoc. Nashville, TN: Hanley & Belfus, Inc., 1997, p. 51–55.

[8] Federal Register, 2000 - HIPAA Privacy Rule – US Code of Federal Regulations (CFR): 45 CFR, Part 160 and Subparts A and E of Part 164.

[9] K. Benitez K & B. Malin. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc. 2010 Mar; 17(2):169–177.

[10] Loukides, G., Gkoulalas-Divanis, A. & Malin, B. COAT: COnstraint-based anonymization of transactions. Knowl Inf Syst (2011) 28: 251-282

[11] M. Sariyar M & I. Schlunder . Reconsidering anonymization-related concepts and the term "Identification" against the backdrop of the European legal framework. Biopreserv Biobank. 2016 Oct 1; 14(5): 367–374.

[12] S. Garfinkel. De-Identification of Personal Information. National Institute of Standards & Technology (US Department of Commerce), pp. 1-54. NISTIR 8053, 2015

[13] D. Barth-Jones. The "re-identification" of Governor William Weld's medical information: A critical re-examination of health data identification risks and privacy protections, then and now. Pre-publication draft – working paper, pp. 1-19. June 18, 2012.

[14] A. Cavoukian, & K. El Emam. Dispelling the myths surrounding de-identification: anonymization remains a strong tool for protecting privacy (June 2011), pp. 1-19. OIPC Ontario.

[15] Z. Wan Z et al. A game theoretic framework for analyzing re-identification risk. PLoS ONE 10(3): e0120592, pp. 1-24, 2015

[16] A. Narayanan & E Felten, No silver bullet: de-identification still doesn't work, pp. 1-8, July 9, 2014, http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf.

[17] X Wang, D Sontag and F Wang. Unsupervised learning of disease progression models. In KDD'14, August 24‑27, pp. 84-94. http://dx.doi.org/10.1145/2623330.2623754.

[18] F. Saad & V. Mansinghka. Detecting dependencies in sparse, multivariate databases using probabilistic programming and non-parametric Bayes. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54:632-641, 2017