Health Care and Informatics Review Online, 2012, 16(2), pg 24-30, Published online at www.hinz.org.nz
ISSN 1174-3379

Assessing Re-identification Risk of De-identified Health Data in New Zealand

Jayden MacRae, Stacey Dobbie, Dipan Ranchhod
Compass Health
PO Box 27380, Marion Square, Wellington, New Zealand
jayden.macrae. stacey.dobbie, dipan.ranchhod@compasshealth.org.nz

Abstract

Aim: *This paper investigates the relationship between re-identification risk, patient cohort sizes and demographic variables in de-identified New Zealand health data.* ***Method:*** *We assessed six combinations of demographic variables (quasi-identifier models) over cohorts ranging from 1,000 to 430,000 patients using patient data from three Primary Health Organisations. We applied a k-anonymous model with a threshold of k=1 to determine uniqueness and re-identification risk.* ***Results:*** *Four of the six quasi-identifier models we investigated exhibited significant risk of re-identification in all cohort sizes. Two quasi-identifier models using the most aggregated quasi-identifiers adequately prevented re-identification when applied to cohorts over 150,000 patients.* ***Conclusion:*** *To minimise re-identification risk, published data sets should be aggregated from more than 150,000 individuals; not include meshblock; and represent age and ethnicities within groups. The best way for trusted health agencies to protect shared micro-data is to ensure data use protocols and agreements are in place.*

1. Introduction

Health data is routinely used for statistical and research purposes in de-identified forms in New Zealand. The New Zealand Health Information Privacy Code allows the use of heath information for research or statistical purposes as long as the information will not be published in a form that can be reasonably used to identify an individual [1]. Typically the way in which health information is de-identified is to remove the National Health Index number (NHI), name and address information. Demographic data increases the utility of clinical data and therefore is important for analysis. The demographic fields often used in health analysis in New Zealand include ethnicity, gender, date of birth or age and deprivation quintile. Meshblock information may sometimes also be included for geospatial analysis.

Previous studies based on other countries' populations have suggested that common demographic information can be used to re-identify individuals in de-identified data sets. El Emam et al describe demographic fields that can be used to re-identify patients as quasi-identifiers and a group of quasi-identifiers as a quasi-identifier model [2]. By identifying unique combinations of quasi-identifier values within quasi-identifier models, de-identified data can be matched back to individuals. This effect has been demonstrated by Golle within the United States of America [3]. Golle calculated that using only gender, full date of birth and postal code, 63% of the population could be uniquely identified. A prior study had calculated an even higher rate of 87% [4].

Countries around the world have different guidelines and protocols in place for those responsible for protecting patient privacy within health information. The Department of Health and Human Services in the United States of America outline 18 demographic fields that should be removed from data used in research to protect personal privacy [5]. The US Bureau of the Census restricts public release of geographic area population size to 100,000 records or more [6]. Many protocols for protecting against re-identification have been suggested [7–13].

This paper investigates if there may be any relationship in the risk of re-identification using New Zealand health data with various population sizes and quasi-identifier models. We hope this paper may enable analysts and researchers to understand the implications of using various quasi-identifier models and sample size selection to prevent re-identification when publishing or showing New Zealand health data.

2. Method

Patient data from three Primary Health Organisations (PHOs) for January to March 2011 was used as the basis for all calculations. Each PHO had funded populations of 40,909, 152,228 and 241,707 totalling 434,844 which represent the four largest cohort sizes in the analysis. It is usual for statistical and research purposes for much smaller populations of patient data to be used, such as in services management or condition specific research reports. To assess risks in smaller cohort sizes such as these, patient data was picked at random from the patient population, forming cohorts of 1,000, 5,000 and 10,000 patients respectively.

Nine quasi-identifiers were used and grouped into six quasi-identifier models. Table 1 shows the composition of each quasi-identifier model. Age was calculated as at 1 January 2011. Ethnicity data was recorded using level two codes (containing 27 distinct codes). The patient's primary ethnicity was used and secondary and tertiary ethnicities were disregarded. Ethnicity groupings were determined on the basis of mapping level two codes to Maori, Pacific, Asian, European and Other for the five category group, and Maori, Pacific and Other for the three category group.

Table 1 - Quasi-identifier model composition and individuals with unique quasi-identifier values by cohort size

		Quasi-identifier Model					
		1	2	3	4	5	6
Quasi –identifiers (Qi)	DoF						
Date of Birth	40,178	✓		✓			
Age	111		✓		✓		
Age Group	10					✓	✓
Gender	4	✓	✓	✓	✓	✓	✓
Ethnicity	27	✓	✓	✓	✓		
Ethnic Group (5)	5					✓	
Ethnic Group (3)	3						✓
Meshblock	46,263	✓	✓				
Deprivation	6			✓	✓	✓	✓
Cohort Size		Rate of People with Unique Qi Values / 100,000 people					
1,000		100,000	99,800	98,200	62,700	10,400	3,900
5,000		99,900	99,100	94,600	27,500	2,200	300
10,000		99,700	98,200	92,700	19,200	900	110
40,909		98,499	92,999	61,099	4,999	90	12
152,228		98,100	79,300	62,700	2,700	0	0
241,707		98,900	69,600	66,200	1,700	0	0
434,844		97,400	57,900	64,000	10	0	0
Cohort Size		Count of People with Unique Qi Values					
1,000		1,000	998	982	627	104	39
5,000		4,995	4,955	4,730	1,375	110	15
10,000		9,970	9,820	9,270	1,920	90	11
40,909		40,295	38,045	24,995	2,045	37	5
152,228		149,336	120,717	95,447	4,110	0	0
241,707		239,048	168,228	160,010	4,109	0	0
434,844		423,538	251,775	278,300	43	0	0

The quasi-identifier models were structured to emulate models commonly used in current practice. Quasi-identifier models one to three represent micro-data, available in various collections such national pharmaceutical and laboratory warehouses, and that may be shared with researchers. Quasi-identifier models four to six represent aggregated data used for service management and contractual reporting purposes. Some quasi-identifiers are related and mutually exclusive. 'Age' and 'age group' represent an aggregation of 'date of birth', the two ethnic groups are more aggregated data of 'ethnicity' and 'deprivation' is derived directly from meshblock.

One-thousand random samples were generated for each quasi-identifier model and the three smallest cohort sizes. Any patient could appear only once in any one sample but was not limited to the number of samples they could appear within. A patient was deemed to be re-identifiable if they were the only patient that existed for any given set of quasi-identifier values. This is described in literature as a k-anonymous threshold of k=1 [7,13,14]. We calculated for each sample the proportion of patients that could be re-identified at this threshold. Each quasi-identifier model had a mean and 95% confidence interval calculated for each of these cohort sizes. The sample sizes were of sufficient size to assume a normal distribution of the data to calculate confidence intervals [15]. Confidence intervals were determined using the values 1.96 standard deviations from the mean.

The same analysis was applied to each of four largest cohorts. Because the analysis was based on each PHOs entire funded population sampling was not undertaken.

The degrees of freedom for each quasi-identifier model were calculated based on a living population aged between 0 and 110, and covered all meshblocks throughout the country.

3. Results

Table 1 contains a summary of how each quasi-identifier model was composed, along with the degrees of freedom (DoF) for each quasi-identifier, and the rate and count of re-identified people for each cohort size. The majority of cohort sizes and quasi-identifier models allow re-identification of individuals. Cohort sizes over 150,000 with quasi-identifier models five and six have no individuals with unique quasi-identifier values. Only this combination of cohort size and quasi-identifier model prevent re-identification based on quasi-identifier values.

As both the quasi-identifiers become more aggregated and the cohort size increases the re-identification rate decreases. This is illustrated in Figure 1. There is an anomaly in this observation in quasi-identifiers one and three within the 152,228 and 241,707 cohorts. Re-identification rate increases from the smaller to larger cohort in both cases.

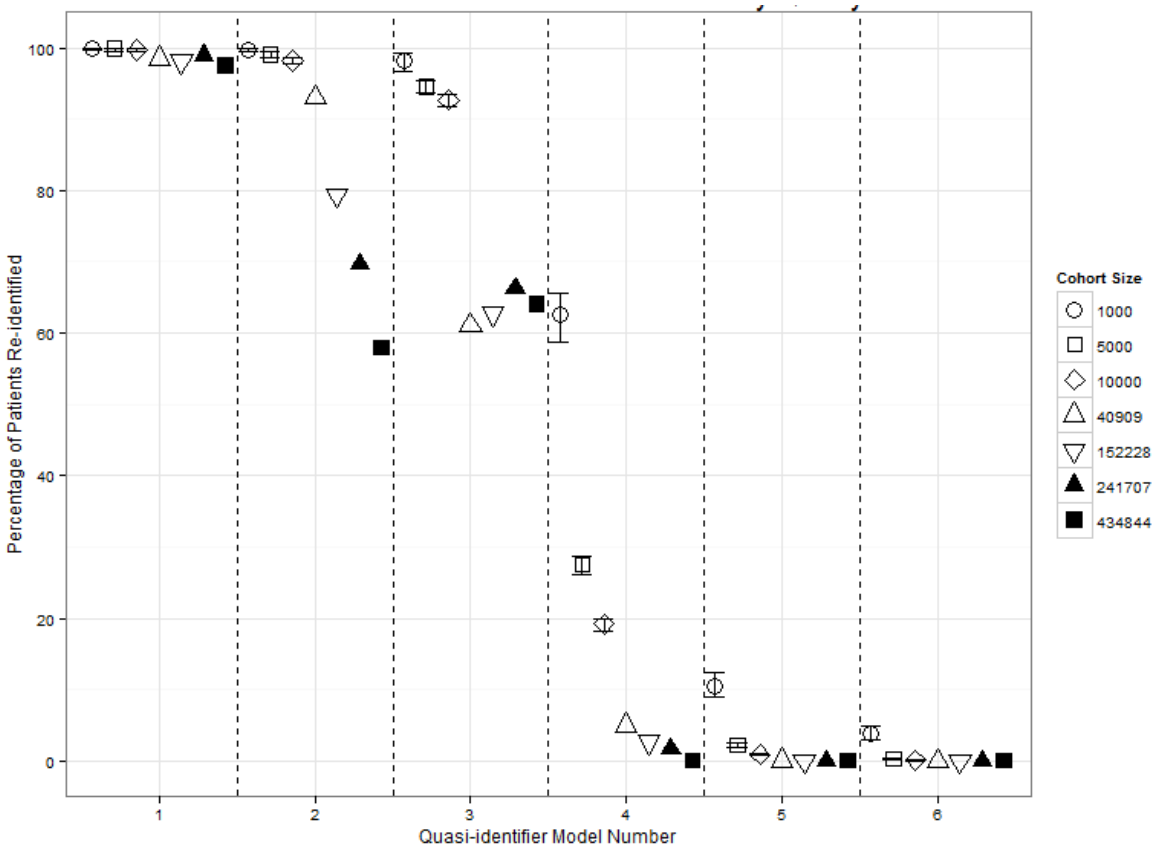

Figure 1 - Re-identification rates for quasi-identifier models and cohort sizes with 95% confidence intervals

The re-identification rate and cohort size when plotted on a log-log axis for quasi-identifier models four to six exhibit a roughly linear relationship shown in Figure 2. The lines of best fit are calculated using nonlinear least squares regression for a power model. Zeros are omitted as log 0 is undefined. The pseudo-R^2 for these fits are 0.9927, 0.9923 and 1.0000 respectively.

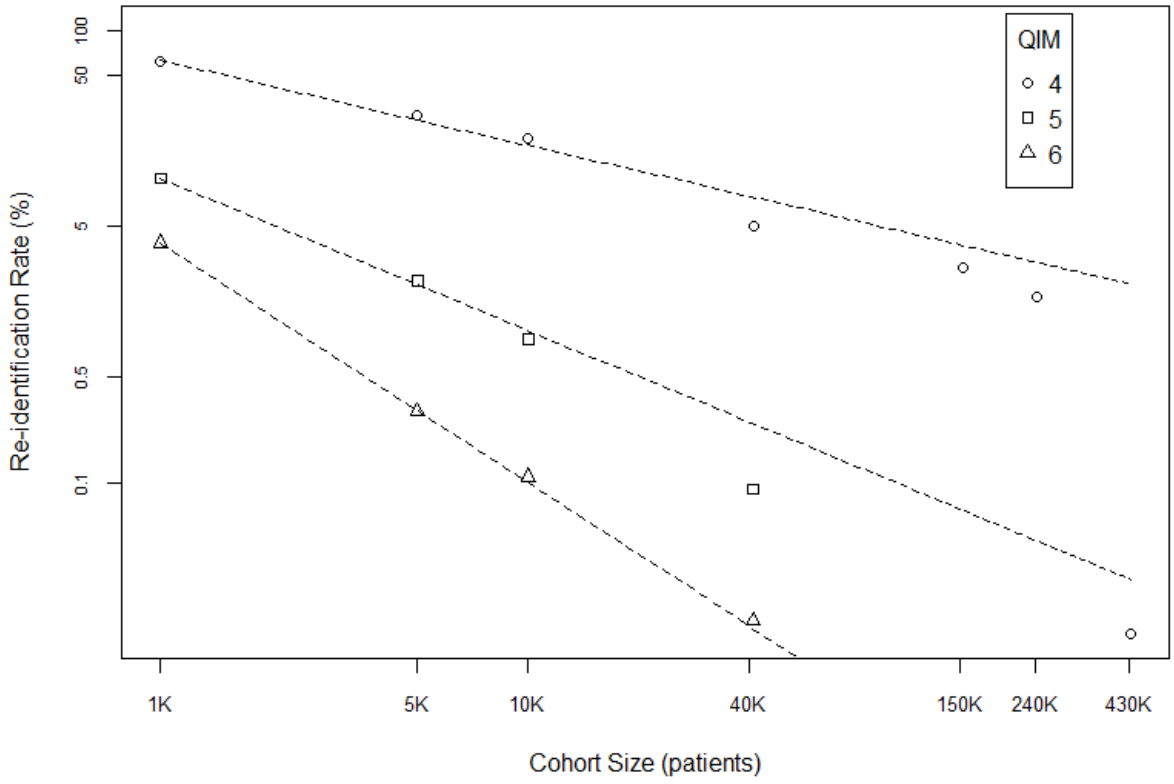

Figure 2 - Percentage re-identified patients for cohort sizes for quasi-identifier models with best fit lines for each model

4. Discussion

None of the quasi-identifier models tested protected against re-identification using a k-anonymous threshold where k=1 for cohort sizes below 150,000 patients. For cohort sizes over 150,000 patients, only the two most aggregated quasi-identifier models prevented re-identification. This would suggest that any datasets that are published below 150,000 patients, or any that don't use the most aggregated quasi-identifiers may carry the risk that individuals may be re-identified, regardless of the model or cohort size used.

New Zealand health data appears to exhibit similar properties of re-identification as seen in other countries. Cohort size thresholds to prevent re-identification have been proposed within other countries, ranging from 100,000 in the US to 70,000 in Canada. [6,16]. The cohort size threshold for preventing re-identification in our data may be between 40,000 and 150,000. Our method had a large break between cohort sizes. Because of this our results can only suggest a threshold of 150,000 patients to prevent re-identification. This is higher than thresholds proposed on analysis of other countries populations. Analysis of the re-identification rates with cohorts between 40,000 and 150,000 may reveal lower thresholds that are more consistent with these. The omission of cohort sizes that may have provided a more precise threshold within this range occurred because the original investigations that produced the data for this paper were to answer questions about research involving sample sizes between 1,000 and 10,000 patients and the PHO populations were included as reference measures. Further research with the goal of determining a more precise population threshold should be completed in the future.

The total population for which we performed the analysis is reasonably representative of the make-up of the overall New Zealand population. The analysis that we performed was on data that was sourced from a geographic region

covering roughly 10% of the New Zealand population. The population proportions for ethnicity, age groups and gender all differ by less than 3% compared with 2006 census data. This suggests that our results are likely to be able to be generalised to other regions within the country.

There appears to be an anomaly in the cohort size and re-identification risk relationship in quasi-identifiers one and three, where the cohort size of 241,707 patients have a higher proportion that can be re-identified than the smaller 152,228 population. This may be a result of more homogenous quasi-identifier values in one population over another. These two quasi-identifier models both contain date of birth. All other models contain age. One explanation for this observation could be that a larger population that is more diverse in age will have fewer individuals that share the same date of birth. While this may affect the generalizability of our results, the magnitude of the effect observed in our data appears to be inconsequential.

The two quasi-identifiers which appear to significantly affect re-identification rate are date of birth and meshblock. Meshblocks are geographical areas of varying size with populations typically between 70 – 110 people. They are the smallest geographic unit that census data is aggregated to, and are used to define boundaries for local and health authorities, as well as being associated with socio-economic deprivation measures. Date of birth and meshblock both have degrees of freedom of 40,178 and 46,263 respectively, much higher compared with other quasi-identifiers (see Table 1). People are not evenly distributed throughout the range of values each of these variables may take, as the majority of people reside in roughly 10% of the possible meshblocks, and there are significantly more people being born today (and sharing birth dates) than there were 50 or 100 years previously. This results in a diminishing of the impact of these degrees of freedom.

The threshold we used to define uniqueness was relatively low. Using a k-anonymous threshold of k=1 means that any combination of quasi-identifier values within a quasi-identifier model that has a count of two or more is deemed to be non-unique and therefore not re-identifiable. Machanavajjhala et al. present an argument that techniques with relatively low sophistication can thwart such low thresholds. [13] Raising the threshold to k>1 will increase the re-identification rate but to what degree is unclear. We believe it is likely that the populations below 250,000 people will exhibit unacceptably high re-identification rates even where k=5. Further investigation using a higher threshold of k=5 or above may be appropriate.

The practical re-identification risk of publically released data is likely to be lower than what we have calculated here without any reliable reference frame. Census data is the most obvious reference frame that could be used for re-identification matching. It is not currently released to the public at a level in New Zealand that we believe would be useful for this purpose. PHO register data is another possible reference frame that could be used for re-identification of patients. This data is not publically available.

Re-identification risk still exists even without a current reference frame however. There are techniques with low sophistication that can be used to re-identify individuals or elicit information on known individuals from health data sets. [9] Our results still suggest caution to those responsible for the release of de-identified and aggregated health data.

Researchers have previously suggested that when sharing health data with trusted agencies or researchers with appropriate security and privacy protocols in place the threshold for re-identification can be relaxed to allow up to 20% re-identification.[2,12,17] Following these guidelines our results suggest that quasi-identifier models one to three should never be used regardless of cohort size, and quasi-identifier model four should only be used in cohort sizes above 10,000 patients. Quasi-identifier models five and six can be used on smaller cohort sizes below 1,000 patients.

The utility of the data offered in quasi-identifier models four, five and six may not be sufficient for the needs of specific research or statistical reporting. In such instances it may be necessary to trust other agencies with de-identified data at re-identification rates higher than 20%, or to offer in-house analysis services to provide the utility on the data sets, but provide aggregated results back to researchers or report recipients. Another alternative could be to provide access to micro-data labs where aggregated results can be removed, but micro-data remains within source organisations.

Publically released health data that is categorised by or limited to PHO or District Health Board (DHB) and contains demographic breakdowns has some risk of re-identification. Nine of New Zealand's 20 DHBs and 27 of 36 PHOs have total populations below the 150,000 threshold. Analysts within these organisations should therefore take care with what information they release that may be re-identified.

The current process of de-identifying national pharmaceutical and laboratory databases serves little function to protect patient identity. These databases provide important clinical and service management information to health agencies such as PHOs and DHBs. They contain detailed information on pharmaceutical dispensing and laboratory testing at a de-identified patient level. The demographics available include age, and level two ethnicity codes, similar to quasi-identifier models one to three. Health agencies that have access to this information have reference frames in the form of patient registers. Our results confirm that they would be able to re-identify the majority of individuals with trivial effort if tey wished.

Having good protocols and agreements around the sharing of data between trusted agencies is more useful than removing or encrypting unique identifiers. When data sets include quasi-identifiers of date of birth, meshblock or even age there is such a high re-identification rate, de-identification serves little function. Removing personally identifying information such as name and address may still be valid to prevent the casual recognition of patients among data sets. Protocols and agreements can act to outline how health agencies will use and protect shared data.

There appears to be a roughly linear relationship between cohort size and re-identification rate when plotted on a double log axis for quasi-identifier models four to six. This may infer a power law relationship between the two variables. Establishing whether empirical data obey a power law is complex because of three factors; assumptions for the calculations of standard errors are not met; the use of r^2 to determine goodness of fit in such a distribution such as a power law has a low statistical power; and our data has a small number of empirical samples to base observations upon [17]. We are cautious therefore to suggest these variables obey such a law. We suggest that future study should be made of a data set with significantly more observations and specific techniques to establish whether a power law distribution exists. If such a law exists, it would be useful for predicting re-identification thresholds for specific cohort sizes.

5. Conclusion

De-identified data sets being shared or published should contain more than 150,000 individuals, not have accompanying meshblock data and should ensure age and ethnicity demographics are aggregated.

When sharing aggregated clinical data with trusted organisations the minimum cohort size will depend greatly on the quasi-identifiers used. For quasi-identifier model four, cohort sizes should be more than 10,000 patients, while for models five and six, they should be more than 1,000 patients.

It is important to have good agreements in place with trusted organisations when sharing micro-data. De-identified data can be re-identified so readily in micro-data quasi-identifier models that it serves little practical purpose to de-identify it initially.

6. Acknowledgments

The authors thank Dr Tom Love from Sapere Research Group and Mr John Grant from Compass Health for providing comment and feedback on the paper.

7. References

[1] Office of the Privacy Commissioner. Health Information Privacy Code 1994. 2008.

[2] El Emam K, Brown A, Abdel Malik P, Neisa A, Walker M, Bottomley J, et al. A method for managing re-identification risk from small geographic areas in Canada. BMC Med Inform Decis Mak. 2010;10(18).

[3] Golle P. Revisiting the uniqueness of simple demographics in the US population. Proceedings of the 5th ACM workshop on Privacy in electronic society. AC, Alexandria, Virginia, USA; 2006;77–80.

[4] Sweeney L. Uniqueness of Simple Demographics in the U.S. Population. Carnegie Mellon University, Laboratory for International Data Privacy; 2000.

[5] Downey P. Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule. 2004.

[6] Hawala S. Microdata disclosure protection research and experiences at the US Census Bureau: An Update. Proceedings of the Workshop on Microdata, 2003. 2003.

[7] Samarati P. Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering. 2001;13(6).

[8] Sweeney L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. 2002;10(5):557–70.

[9] El Emam K, Buckeridge D, Ramblyn R, Neisa A, Jonker E, Verma A. The re-identification risk of Canadians from longitudinal demographics. BMC Med Inform Decis Mak. 2011;11.

[10] El Emam K, Dankar F. Protecting privacy using k-anonymity. J Am Med Inform Assoc. 2008;15(5):627–37.

[11] Howe HL, Andrew JL, Shen T. Method to assess identifiability in electronic data files. J Am Med Inform Assoc. 2007;165(5):597–601.

[12] LeFevre K, DeWitt D, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity. Proceedings of the 2005 ACM SIGMOD international conference on Management of data. New York, NY, USA: ACM; 2005. p. 49–60.

[13] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD). 2007 Mar;1(1).

[14] Statistics Canada. Canadian Community Health Survey Cycle 3.1 Public Use Microdata File User Guide. 2006;

[15] Chang H, Huang K, Wu C. Determination of Sample Size in Using Central Limit Theorem for Weibull Distribution. Information and Management Sciences. 2006;17(3):31–46.

[16] Howe H, Lake A, Lehnherr M, Roney D. Unique record identification on public use files as tested on the 1994-1998 CINA analytic file. North American Association of Central Center Registries. 2002;

[17] Clauset A, Shalizi CR, Newman MEJ. Power-law Distribution in Empirical Data. Society for Industrial and Applied Mathematics Review. 2009;51:661–703.