

Using Census Records to Advance Genetic Research

Terry Quinlan, Atlantic Coordinator

Canadian Century Research Infrastructure

Contact: tquinlan@gov.nl.ca

In collaboration with:

Newfoundland and Labrador Statistics Agency, and

Population Therapeutics Research Group, Memorial University

International Sociological Association Research Committee on Family Studies

Health in Families, Healthy Families: Gendered Explorations

Toronto, Canada

May 8-11, 2007

Session: Canadian Censuses and Family Research

Organizer: Dr. Peter Baskerville

Chair: Dr. Gordon Darroch

Abstract

Currently, health researchers looking for genetic links for a particular disease spend years visiting residents in small communities and searching through church records and other archives. This process is time consuming, expensive, and often incomplete. Over the years, various computer models have been developed for data mining and record linkage to expedite this process. In this report, we demonstrate the effectiveness of an iterative learning system for reconstructing family tree information over large populations. We make the case for using manuscript census records as the primary data source for attaining optimal results, and we demonstrate the importance of leveraging the wealth of household and neighbourhood contextual information available in the census to achieve this goal.

Acknowledgements

The Canadian Century Research Infrastructure (CCRI) project is a five-year pan-Canadian initiative to develop manuscript records for the 1911 to 1951 decennial censuses for historic, social, and related research. This includes, for the first time, integrating corresponding records from the pre-Confederation Newfoundland and Labrador censuses for 1911, 1921, 1935 and 1945. Partners in the CCRI are the University of Ottawa (lead institution), University of Victoria, University of Toronto, York University, Université du Québec à Trois-Rivières, Université Laval, and Memorial University. Memorial University, in collaboration with the Newfoundland and Labrador Statistics Agency (NLSA), is responsible for developing the Atlantic Canada portion of this database, including pre-Confederation Newfoundland and Labrador. Major funding for this project was provided by the Canada Foundation for Innovation.

The construction of family tree information for genetic research from the Newfoundland and Labrador portion of the data is the first major adaptation of the CCRI infrastructure. Partners with CCRI in this endeavour are NLSA and the Population Therapeutics Research Group (PTRG), a non-profit research organization within the Faculty of Medicine at Memorial University, as well as the Provincial

Archives, Rooms Corporation, Government of Newfoundland and Labrador, stewards of the pre-Confederation historic records.

Introduction

Newfoundland and Labrador's mainly English and Irish population is composed primarily of descendants from a limited founder population that settled here in the 1700s. Over time, geographic isolation and limited migration has led to a higher incidence of certain diseases, such as juvenile type 1 diabetes mellitus, colorectal cancer and psoriatic arthritis¹. For this reason, research in this province can contribute tremendous insight into the genetic basis for disease and supplement the quest for best treatment options.

As a result, there is considerable interest among the medical community in doing research here. Historically, various sources from pre-Confederation Newfoundland (such as manuscript census, vital statistics, and parish records) have been available to researchers, historians, and the public at large, making them immensely important for detailed investigation of this nature. However, in the absence of critical computerized databases, health researchers looking for genetic links for specific diseases often spend years searching through these various archives. To aid in the search, researchers will often supplement their findings by visiting local area residents to gain additional insight into family relationships. This process is by its very nature expensive and time consuming and often yields incomplete results.

Construction of a centralized family tree database will help facilitate research into topics such as the genetic factors contributing to disease and complement current work in the field of pharmacogenomics. Pharmacogenomic research is a relatively new area of medicine that can lead to more effective prescribing based on individual genetic profiles – the right drug for the right patient at the right dose². Population based research provides the opportunity to study a group of people on the same medication who share a similar genetic background. This can make it easier to pinpoint genetic

similarities that may have an impact on the effectiveness of the medication. As outlined earlier, Newfoundland is an especially good population for these types of studies.

For this reason, CCRI, in partnership with NLSA, has been investigating ways to develop family tree information on a province-wide scale in a cost effective and timely manner by using both manuscript census and other records. After an initial cost-benefit assessment, it was determined that of the sources available, the Newfoundland and Labrador pre-Confederation census records were the one source that could allow the most linkage success at the lowest incremental cost, so we began our further testing and development there. Over time, other data resources will be leveraged as required to complete as much as possible of the entire pedigree.

Once constructed, the family trees will be transferred to PTRG, who will combine this resource with patient information to develop a province-wide heritability database. This linked infrastructure can then be made available to health researchers for genetic research.

Of course, medical ethics are a primary concern once personal medical information is used in any way. We recognize that privacy and protection of the rights of the individual are paramount. For this reason, all data used by PTRG is stripped of identifying information and guarded by strict privacy policies based on federal and provincial legislation. Each project that uses the database is reviewed by Memorial University's Human Investigation Committee to ensure the highest ethical standard is upheld.

By allowing researchers to find critical genealogical links in minutes instead of years in a way that better protects individual privacy, the conclusion is that this approach has the potential to dramatically change the way genetic analysis is conducted.

In this report, we make the case for using manuscript census records as the primary data source for reconstructing family tree information over large populations, with other data to follow in a progressive, stepwise method. We illustrate the effectiveness of an iterative, learning system for this purpose and highlight the importance of leveraging the wealth of household and neighbourhood contextual information available in the census for achieving optimal results.

Related Research

The large-scale use of census and other historic records to reconstruct family history for genetic and other research is not new. For example, the BALSAC register of census and other records in Quebec is widely used for studies of human genetics. It began in 1972 at the Université du Québec à Chicoutimi based on the Saguenay region. Since then, the partnership has expanded to include Université Laval, McGill, and Université du Montréal, with the goal of extending the register to the entire population of Quebec³. Similar work has been reported in numerous other countries.

The Census as the Primary Source

In our initial assessment, we considered the benefit of various sources available for our study region. Aside from manuscript census records, these included vital statistics for births, marriages and deaths; parish records of baptisms, marriages and burials; and other sources such as business directories, voters lists and transcriptions of cemetery headstones.

We found the rate of success in reconstructing entire families was considerably higher using the census than with any other source we investigated. This is mainly due to three reasons:

- 1 – Due to its nature, the manuscript census already joins together many of the members of most households that existed at the time the census was conducted, so the number of additional links necessary to complete a family is reduced.

2 – Information from these individuals then provides additional context to ensure a higher level of confidence in identifying any other family links. This is especially true for frequently used first names where substitution of like individuals is a concern.

3 – Non-household neighbourhood information also contributes significantly to match rates. This is illustrated later in this discussion.

The Effectiveness of an Iterative Learning System

The mathematical model pioneered by Ivan P. Fellegi and Alan B. Sunter⁴ still provides a framework for many of the models currently used to match individuals across data files.

CCRI has modified this approach by including household and neighbourhood information. We recognize that this would potentially introduce additional selection bias, such as increasing the likelihood of matching those who stayed in the same household over those who did not. However, our goal is to develop a complete as possible dataset for genetic, as opposed to historic, research, so bias is less of a concern for us than coverage.

More importantly, most existing methodologies are designed to match on information from the individual alone. However, because we are interested in leveraging any relevant and available contextual information, we felt that using these approaches would not produce the best results for our purpose.

After considering various options, we decided to evaluate an iterative learning approach instead. To do this, we needed to create a database for testing and development and chose to examine Notre Dame Bay South, a rural area where the author has family connections and could rely on other resources to help identify linkages as required. In total, 11,527 records from the 1935 census were selected for further exploration. With our local research, we were then able to match 4,329 of these individuals

manually with their corresponding records in the 1921 census. These linked record pairs became the database we used to build our model.

Unmatched Individuals

The following is an examination of the 7,198 selected individuals who were not matched in our database and not included in our analysis.

Born since previous census	53.2%
Marriage	11.2%
Other	35.6%
Total	100.0%

The single biggest reason for a non-match is the individual was simply not yet born at the time of the previous census and so no link was possible (53.2% of non-matched cases). Another 11.2% of those not linked was the result of name changes due to marriage. This highlights the need to use various marriage records as an important secondary source to provide additional required information, specifically on maiden names.

The third group, comprising 35.6% of non-matches, is mainly due to incomplete original manuscript census records from areas neighbouring our database. Of these, the missing records for the Fogo district have the largest impact because of the geographic proximity and strong family connections between the two areas. Based on research into other sources, many of the non-matched individuals in the pilot data are known to have been residing in the Fogo district at the time of the 1921 census. Our rates of non-coverage are consistent with expected migration flows between these regions.

This lack of complete information is a problem with the census and other historic sources, creating special challenges for work of this nature. To address this issue, we need to supplement census

records with other sources in areas where this is the case. For example, in the Fogo district, while there are no census records for 1921, there is instead relatively good coverage from parish and other records.

Building the Model

The 4,329 linked pairs we manually identified became our learning database for testing and improving our automated match rates by:

- 1 – Computing the preliminary model with initial parameters and weights and identifying clusters of non-matches
- 2 – Investigating reasons for these non-matches
- 3 – Refining and adjusting the system accordingly

The above steps were then repeated on an iterative basis until various patterns of non-matches were found and the various weights were optimized to produce maximum match results within the learning database. Intuitively, one expects that the lessons learned and the success rates with such a sample database can then be extrapolated to the overall population.

One disadvantage of the current database is that it is restricted to one geographic area and, in some ways, may not accurately reflect the entire population. For this reason, the database will eventually be extended to include one additional rural and one urban area, as one expects some differences due to urban-rural factors. The model may need to be adjusted somewhat at that time to reflect this reality.

The Importance of Leveraging Neighbourhood Context

In testing the learning database, we recognized the potential contribution of household and neighbourhood contextual information to match results.

Therefore, to measure the impact of context, we ran the same model under three scenarios:

- 1 – Individual information only
- 2 – Contextual information from household members
- 3 – Contextual information from all neighbours

The results are as follows:

<u>Context</u>	<u>True</u>	<u>False</u>	<u>None</u>	<u>Accuracy</u>	<u>Coverage</u>
Individual	87.6%	7.0%	5.4%	92.6%	94.6%
Household	93.5%	2.7%	3.8%	97.2%	96.2%
Neighbour	97.1%	1.5%	1.4%	98.5%	98.6%

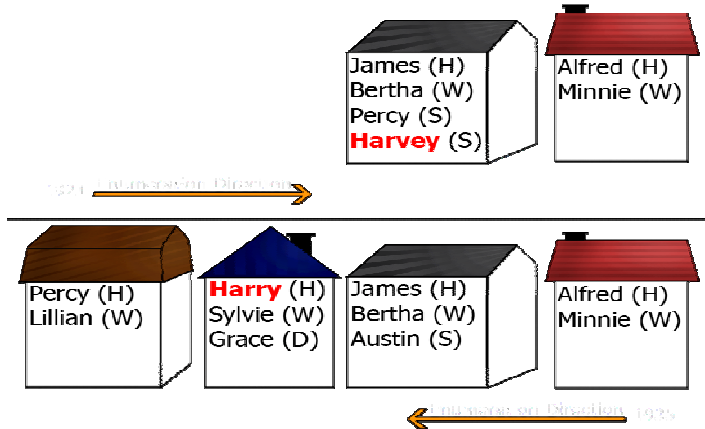
For individual information only, we had a true match on 87.6% of previously identified matched records, a false match on 7.0%, and no match on the remaining 5.4%. We define accuracy as the number of correctly identified individuals over the total matches, and coverage as the number of matches over the total possible. Under this scenario, the accuracy was 92.6% and the coverage was 94.6%.

If we include household contextual information, true matches increase to 93.5%, and both accuracy and coverage increase significantly, to 97.2% and 96.2%, respectively.

Furthermore, if we include additional (non-household) contextual information, true matches now increase further to 97.1%, with corresponding increases in both accuracy and coverage. In other words, there were additional considerable improvements in matches by the inclusion of non-household contextual information.

To understand why this is so, allow us to demonstrate using an actual example from the census records:

In 1921, Harvey was living with his parents, James and Bertha, and his brother, Percy. By 1935, however, he was recorded under the name Harry, had a wife named Sylvie as well as a daughter named Grace, and was now living in his own home next door to his parents. His brother Percy also had a home nearby. Neighbours Alfred and Minnie remained in the same location.



Based on individual information alone, we did not recognize that Harvey and Harry was the same person, mostly because of the name change.

Similarly, no match was found based on household contextual information. This is because the specific household from 1935 (Harry, Sylvie and Grace) did not exist in 1921.

Only when we include non-household contextual information does the model produce a true match. It recognized the similarity in non-household individuals (in this case brother Percy now living on one side, parents James and Bertha living on the other side, and neighbours Alfred and Minnie remaining next door to the parents).

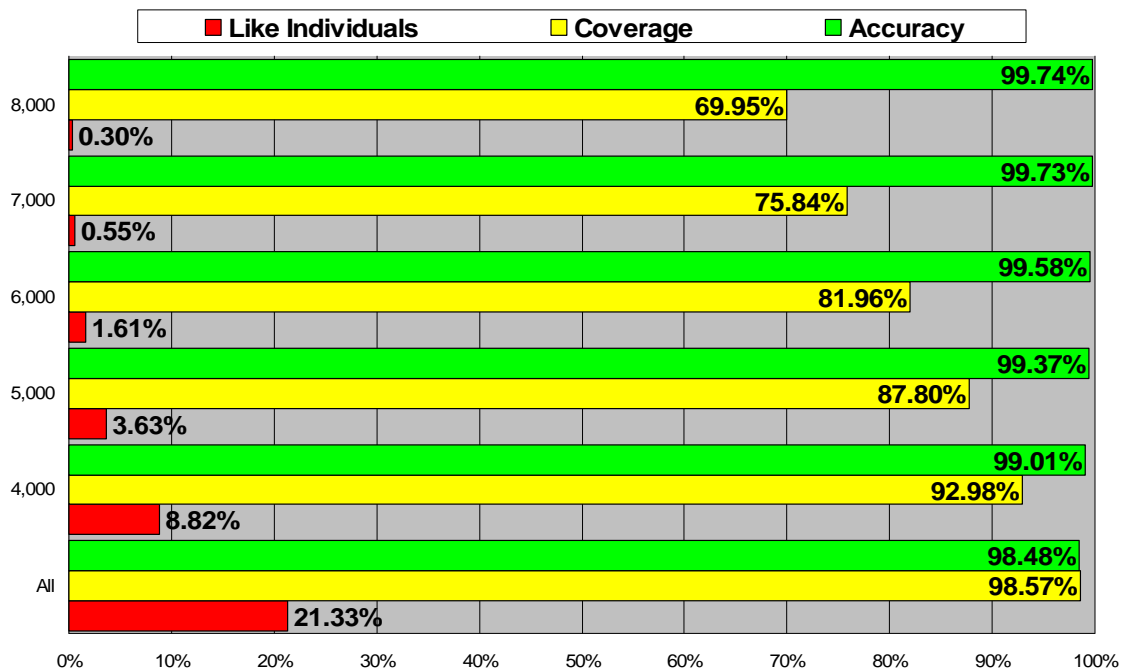
Impact of Thresholds

Threshold values computed by summing all of the weight results for a specific linked pair can be used to classify potential matches into three groups:

- 1 – Above the upper threshold. These are automatically accepted.
- 2 – Below the lower threshold. These are automatically rejected.

3 - Between the upper and lower thresholds. These are set aside for manual inspection.

Decisions on upper and lower threshold have cost-benefit implications. For example, in the following chart we see that for all linked pairs we have an accuracy of 98.48% and coverage of 98.57% for the learning database. To improve accuracy beyond this point requires a corresponding and significant decrease in coverage. If we accepted only cases for which the threshold value exceeded 8,000, say, we can achieve a 99.74% accuracy rate. However, coverage would then be reduced to 69.95% of possible matches.



Furthermore, the issue of like individuals is a concern. While our model has demonstrated success in identifying the correct match in cases where the individual is actually in the database in both 1935 and 1921, substitution of like individuals may become an issue when this is not the case. For example, a record may be unavailable for both years because of missing census records (such as for Fogo) or simply because the person migrated into Newfoundland and Labrador during the intercensal period. Since we are linking from 1935 back to 1921, loss of possible matches due to out-migration and mortality are not considered major factors in this situation.

To demonstrate the magnitude of this potential problem, we removed all known matched individuals and their respective household members from the dataset for 1921 and re-ran the model to assess the probability of false positives. Our analysis indicates that of all records, we then incorrectly matched to another like individual in 21.33% of cases. As evident from the previous chart, the selection of appropriate upper and lower thresholds is critical to reducing this problem to a reasonable level.

Conclusion

The use of historic census records as a first source reduces the number of links required to reconstruct families for the purpose of genetic research. Furthermore, the inclusion of census neighbourhood contextual information significantly improves the ability to link individuals and families from one census to the next. As is the case in any probabilistic match, though, the choice of upper and lower thresholds has an impact on accuracy and coverage. Choosing the appropriate thresholds is also important to reducing the possibility of substitution by like individuals, particularly where missing records are a concern.

¹ Proton Rahman, Albert Jones, Joseph Curtis, Sylvia Bartlett, Lynette Peddle, Bridget A. Fernandez and Nelson B. Freimer (2003), "The Newfoundland population: a unique resource for genetic investigation of complex diseases", *Human Molecular Genetics*, vol. 12, no. 2, pp. 167-172.

² Hong-Guang Xie and Felix W. Frueh (2005), "Pharmacogenomics steps toward personalized medicine", *Personalized Medicine*, vol. 2, no. 4, pp. 325-337.

³ Gérard Bouchard, Raymond Roy, Bernard Casgrain and Michel Hubert (1995), "Computer in Human Sciences: From Family Reconstitution to Population Reconstruction", in Ephraim Nissan & Klaus M. Schmidt (eds), *From Information to Knowledge*, pp. 201-226.

⁴ Ivan P. Fellegi and Alan B. Sunter (1969), "A Theory for Record Linkage", *Journal of the American Statistical Association*, vol. 624, no. 328, pp. 1183-1210.