Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

The challenges of producing national estimates of child maltreatment using administrative data from different jurisdictions

David Laferrière, and Catherine Deshaies-Moreault[1]

Abstract

Statistics Canada was approached to conduct a feasibility study on how to develop a surveillance system for child maltreatment. The Canadian Reported Child Maltreatment Surveillance System (CRCMSS) would integrate data from child welfare agencies in each province and territory to calculate annual estimates of child maltreatment in five categories: physical abuse, emotional abuse, sexual abuse, neglect, and exposure to intimate-partner violence. To reduce burden on child welfare workers, the primary source of data would be a census of administrative data.
We discuss the challenges that must be overcome in order to implement CRCMSS, including the fact that each jurisdiction has its own legislation that defines and categorizes child maltreatment, as well as different systems to identify and track cases of maltreatment, which means the contents and structure of their administrative data and associated text narratives can vary significantly.

For CRCMSS, machine learning techniques from natural language processing will be explored to determine if cases of maltreatment could automatically be identified and classified from narrative reports and an existing dataset. We discuss the practical and technical challenges of using traditional approaches as well as more modern techniques to create coherent national estimates from the administrative data of 13 sub-national jurisdictions.

Key words: Child Maltreatment; Machine Learning; Administrative Data.

1. Introduction

1.1 Background

There is a common interest among governments, non-governmental organizations, physicians, and others in having accurate and timely data in order to help understand the nature (type of maltreatment) and scope (counts and demographic characteristics) of child maltreatment (CM) within the geographic regions they serve (Potter, Hovdestad, & Tonmyr, 2013). CM refers to physical, sexual, or emotional abuse, neglect, or exposure to intimate partner violence of a person under 18 years of age.

In spite of their recognized necessity, the sources of child maltreatment data available in Canada are very limited. Furthermore, each province and territory is responsible for passing and enforcing its own child welfare legislation, which exacerbates the challenges of collecting Canada-level child maltreatment data and using that data to produce estimates.

1.2 Currently available child maltreatment statistics[2]

[1] David Laferrière, Statistics Canada, 100 Tunney's Pasture Driveway, Canada, K1A 0T6 (david.laferriere2@canada.ca); Catherine Deshaies-Moreault, Statistics Canada, 100 Tunney's Pasture Driveway, Canada, K1A 0T6 (catherine.deshaies-moreault@canada.ca).

[2] For a more complete report on the child maltreatment data sources available in Canada, see Potter, Hovdestad and Tonmyr (2013).

Many provinces and territories publish annual reports on child maltreatment, but the data contained in these reports are typically in the form of aggregate tables with little granularity. For example, there is often no breakdown of cases of child maltreatment by type (physical abuse, neglect, etc.) or by age group.

The Canadian Incidence Study of Reported Child Abuse and Neglect (CIS) is a periodic[3] survey that aims to provide a nation-wide profile of the children and families receiving child welfare services. It samples child welfare workers in each province and territory who complete a questionnaire for each of the cases for which they are responsible that were newly opened within a given reference period. Among its primary objectives, CIS was designed to "determine rates of investigated and substantiated physical abuse, sexual abuse, neglect, emotional maltreatment, and exposure to intimate partner violence, as well as multiple forms of maltreatment" (Public Health Agency of Canada, 2010). The CIS is a valuable source of child maltreatment data, but its periodic nature substantially limits its timeliness.

There are other sources of child maltreatment data in Canada, but they currently form a somewhat incomplete patchwork (Potter, Hovdestad, & Tonmyr, 2013) that does not easily allow researchers to estimate timely, nation-wide statistics.

1.3 Feasibility study

In 2017, the Public Health Agency of Canada (PHAC) approached Statistics Canada to conduct a feasibility study on the development of a surveillance system[4] for reports of child maltreatment. This system would be entitled the Canadian Reported Child Maltreatment Surveillance System (CRCMSS). Its primary aim would be to integrate data from child welfare agencies in each province and territory to calculate annual counts of the number of children investigated for CM and the total number of investigations in five categories: physical abuse, emotional maltreatment, sexual abuse, neglect, and exposure to intimate partner violence (EIPV). These counts would incorporate the demographic characteristics of the families referred to in the data collected.

A principal component of the feasibility study was a series of consultations begun in 2018 between Statistics Canada, PHAC, and the provincial and territorial departments responsible for child welfare, with the aim of completing consultations with each province and territory by early 2019. One of the main goals of these discussions was to determine how each jurisdiction captured and stored the data relating to their child welfare reports and investigations. Another goal was to determine how willing each jurisdiction would be to participate in CRCMSS under a variety of data collection scenarios. These scenarios included a survey of child welfare workers and an administrative census.

After consultations and a review of existing data sources, Statistics Canada prepared a draft feasibility study that compared and contrasted the viability and effectiveness of two options for collecting child maltreatment data: a survey of child welfare workers and an administrative census. Due to the burden on child welfare workers and the cost associated with a survey, the feasibility study recommended a census using the child maltreatment data collected by the child welfare agencies in each province and territory. However, there are several challenges associated with collecting and combining the administrative data of the 13 jurisdictions in Canada.

2. Administrative data challenges

Administrative data represent a great opportunity for National Statistical Offices (NSOs). Indeed, the use of administrative data in official statistics decreases response burden, may offer more timely data, and can improve quality by diminishing or eradicating sampling errors. On the other hand, they come with their own challenges. By definition, these data are not collected for statistical purposes, and the scope or definition of the information collected might not be totally in line with the surveillance objectives for which the use of administrative data are being examined. The target population from the administrative data and that of the survey might not be exactly the same, either. Furthermore, some pre-processing of the data may be conducted by the organisation collecting the information without

[3] CIS has been conducted in 1998, 2003, and 2008.

[4] The phrase "surveillance system" is used here in the epidemiological sense meaning the collection, analysis, and interpretation for action of health data (Potter, Hovdestad, & Tonmyr, 2013).

the NSO being aware of it. In addition to these usual challenges with administrative data, CRCMSS has its own specific ones: in particular, the data of interest are governed by sub-national legislations and are contained in an unknown number of data systems.

2.1 Sub-national legislation

Each province and territory is responsible for its own child welfare legislation, and the categories of maltreatment defined in a given province might not map directly to those of CRCMSS. We will also have to work with partners in the provinces and territories in order to stay up to date on all changes in child welfare legislation for each jurisdiction, as these changes may have an impact on the data being captured by a given jurisdiction.

2.2 Data systems

The 13 sub-national jurisdictions (10 provinces and 3 territories) of Canada vary in their child maltreatment assessment tools and data systems, and even within a given jurisdiction there may be inconsistencies in use of both tools and systems. Thus, even the combined data for a given province or territory may be heterogeneous. The quality of the information captured (in the context of producing official statistics) may vary both between and within jurisdictions. As a consequence, we are expecting that significant efforts will be necessary to harmonize the data and derive national figures.

Informatics systems are not static, and neither are the child welfare assessment tools, which evolve over time based on changes in the legislation or improvements to the assessment methods. As a consequence, the CRCMSS project will require monitoring of any changes that may impact how the data are captured or how they are submitted, and its systems and processes will have to be updated accordingly.

Finally, we must be mindful of how the two units of interest for CRCMSS (the child and the investigation) are being captured in the databases. For example, jurisdictions may collect data at the child level or at the family level. From one system to another, the capture of instances where the same child is the subject of multiple investigations will likely vary. We will have to carefully evaluate each system to ensure that we are not over-counting either children or investigations.

3. Coding of narratives

Typically, child welfare workers enter data from their cases using an electronic form with some combination of drop-down menus, check boxes, and text fields. These text fields usually include space(s) for long-form text narratives describing the details of a case. As not all information necessary to capture the five CM categories may be available in the drop-down variables from the different systems, we will most likely have to use the information contained in the text narratives accompanying each case to ensure we are not under-estimating reported CM. For example, we will almost certainly have to use the narratives to supplement the EIPV category in some jurisdictions. Deriving the necessary variables from unstructured text narratives will require coding, *i.e.*, from each text narrative, data for each of the five categories of CM will need to be captured. Human coders, auto-coding algorithms, or a combination of both can be used to obtain the necessary data from the narratives. It is relevant at this point to discuss the following aspects of the coding exercise: obtaining “gold standard” data, human coders and machine learning.

3.1 “Gold standard” data

Whether auto-coding, human coding or a combination of both is used to code the narratives, reference labeled data, or “gold standard” data, will be needed to train the coders and/or the algorithms. Ideally, the labeled data would be coded by those individuals most knowledgeable about the data and systems – the child welfare workers. However, before they code the gold standard data, they should also be trained on the five CM categories (and related variables) to ensure that the information is being mapped uniformly across the country.

There are ongoing discussions between Statistics Canada and the Public Health Agency of Canada on what would be the ideal approach to obtaining gold standard data. There are two approaches under consideration: each jurisdiction providing training examples in the five categories, or a centralized group of experts on the national categories coding each jurisdiction's data. One challenge to the first approach is that welfare workers are unlikely to have time to spare coding data. Hence, it would be very difficult to get gold standard data from all jurisdictions coded by currently employed child welfare workers. However, it might be possible to hire former child welfare workers, particularly recent retirees, to do the coding. The second approach also has challenges, the most obvious of which is finding and training people to become experts on both the five categories of CM and the legislation for 13 jurisdictions.

Finally, it will be important to keep in mind that, even if the people creating the gold standard data are very knowledgeable, human mistakes are possible, so the training data will likely contain errors.

3.2 Human coders

Statistics Canada has a department dedicated to coding, including coding done by humans. The usual process for human coding is as follows: first, subject matter specialists with expertise on how the data should be coded develop the training material. Next, using the training material and examples (possibly coming from the gold standard data), the coders learn the coding strategy. They then code data for which the output is already known, and their success rate and consistency coding this reference data are evaluated. Finally, the training and information provided to the coders is adjusted as necessary, and the process repeated, until a given level of satisfaction with the coding is reached.

Developing training materials, training coders, evaluating them, and employing them to code data is very labour intensive and costly. There are also sources of errors that are difficult to mitigate, and this is particularly true when the concepts to be coded are very complicated (as they are for CRCMSS). In particular, it is difficult to ensure that coders are consistent with each other, *i.e.*, that two coders will read the same narrative and code it identically.

In 2016 PHAC conducted a study to examine how effective human coders could be at coding child maltreatment data from administrative data and text narratives (Tonmyr, et al., 2018). In their study, 12 child welfare workers (CWWs) from one province were trained on PHAC's definitions of child maltreatment, and they provided data based on their CM investigations pertaining to 187 children. These data were used as gold standard data. Two coders were then trained on the five categories of CM, and they subsequently coded the same information that the CWWs had submitted to the administrative data system, without having access to the CWWs' first-hand knowledge of the case. The results of the coders were then evaluated by comparing them to the gold standard data. The results were promising: the two coders' classification of physical abuse, sexual abuse, and neglect broadly matched those of the CWWs. However, the coders had difficulties consistently and accurately identifying sexual abuse, and coding of exposure to intimate partner violence could not be evaluated.

Two of the challenges associated with human coders—cost and inconsistency/inaccuracy—can be mitigated somewhat by using automatic coding techniques. It is unlikely that the use of human coders could be avoided altogether, but automatic coding using machine learning techniques could at least reduce the need for manual coding.

3.3 Automatic coding

Automatic coding refers to coding that is done algorithmically using an automated process, as opposed to manual coding which requires decisions to be made by humans. There are several advantages to automatic coding: it is typically faster, cheaper, and more consistent over time. Automatic coding also allows different types of information to be leveraged. In particular, both text narratives and categorical variables can be used by a coding algorithm, which would be a much more time-consuming task for a human coder.

Most automatic coding algorithms (also called auto-coders) currently used by Statistics Canada are rules-based, *i.e.*, they use some form of decision tree or set of predetermined rules to select outcome codes. However, due to the complexity of child maltreatment concepts and the fact that any text that accompanies a record we receive will be of substantial length and complexity, it is almost certain that a rules-based auto-coder will not be sufficient for CRCMSS. Instead, we aim to use machine learning algorithms to code at least some of the records.

3.3.1 Machine learning for event identification/text classification

Significant work has been done using machine learning to interpret text narratives in a variety of different applications, such as coding cause of death from verbal autopsy reports (Danso, Atwell, & Johnson, 2014), movie review sentiment analysis (Zhang, Marshall, & Wallace, 2016), and coding of occupational injury data from workers compensation claims (Measure, 2014). The principal techniques used in these applications have a common foundation: the methods are based on converting the text into a numerical vector where each row of the vector represents a word.

There are many techniques from natural language processing that can aid in this vectorization process, such as first removing stop words (common words that do not add significant meaning) and stemming (representing words with a common base such as "injured" and "injury" by a single stem word "injure"). These techniques reduce the dimension of the vectors representing the narratives while theoretically preserving the narrative's meaning. After the removal of stop words and stemming, the text narrative is converted into a vector where each row represents the number of times a word appears in the text. The gold standard narratives in vector form are the data used to train, test, and validate machine learning algorithms. Once an algorithm is chosen, it is used to classify new data.

A significant benefit of representing text by vectors in this way is that it is straightforward to incorporate any categorical or numerical data that accompanies the text into the vector itself by simply adding the appropriate dimensions to represent that data. For CRCMSS, it is expected that the narratives from many jurisdictions will be accompanied by some categorical and numerical data, which can be incorporated into any algorithms we develop.

There has been a substantial amount of research into the use of machine learning to identify events from the text narratives included in workers compensation claims. Several machine learning algorithms have been successfully implemented in this field, including support vector machines (SVM), Naïve Bayes and Fuzzy Bayes classifiers, and regularized logistic regression (Marucci-Wellman, Corns, & Lehto, 2017). These algorithms have been compared with one another, and Marucci-Wellman *et al.* found that while logistic regression performed better than the other algorithms individually, the best classifier they were able to build actually used Naïve Bayes and SVM in combination. Specifically, a given case would only be classified automatically if both the Naïve Bayes classifier and the SVM classified the case into the same category. Otherwise, it would be referred to human coders for manual coding.

Among the literature we reviewed, there appears to be a consensus that machine learning algorithms have not yet advanced to the point where they can replace human coders entirely, and that the most effective use of these algorithms is to automatically code some cases and refer the more "difficult" cases to manual coders. Most text classification algorithms work by estimating a likelihood that a case falls into a given category, so an obvious way to use such an algorithm is to choose an acceptable threshold for the predicted likelihood, above which a case is coded automatically. For example, one could train a Naïve Bayes classifier that automatically codes all cases where the predicted likelihood of being in a given category is greater than 80% and refers to human coders all cases that fall below that value. The optimal value of this threshold is context-specific, and great care must be taken when choosing it.

Of the existing research on event identification, the work done on workplace injury narratives appears most applicable to our own project, and we will use this research as a springboard for our methods. Indeed, the computer code and data used by many researchers in that field have been made publically available, which has already greatly accelerated our own work.

4. Ongoing and future work

There are two main tasks that comprise the future work to code CM data for use by PHAC and its partners. In the near term, we will continue our work developing a Python pipeline, which we will use to test a variety of algorithms on text narrative data from different sources. In the longer term, ongoing development will be necessary to update the auto-coder as legislation and data systems change and as data become available.

We have used a preliminary version of our Python pipeline to successfully implement regularized logistic regression, support vector machines, and Naïve Bayes algorithms to automatically classify workplace injury narratives that are publically available online. As we receive child maltreatment data from provinces and territories, we will develop an algorithm (or a combination of algorithms) for each of the unique data configurations, rather than a single algorithm for all provinces and territories. If we were to only develop a single Canada-wide algorithm, it is likely that it would only be effective at classifying narratives from the largest provinces, which is obviously not ideal. However, the timeline for receiving the necessary data has not yet been finalized.

Child maltreatment data are incredibly sensitive in nature, and the fact that legislation and data systems vary between the provinces and territories makes the process of data acquisition by Statistics Canada particularly complicated. Furthermore, we will require data that have already been classified by child welfare workers to use as gold standard data for training our algorithms, and it is not yet clear how this work will be done. In the meantime, Statistics Canada is investigating the possibility of acquiring child maltreatment survey data and narratives from other sources that we could use to do preliminary work training algorithms and identifying vocabulary and phrases specific to that subject.

5. Conclusion

There is a clear need for comprehensive nation-wide data on child maltreatment. In order to help meet this need, Statistics Canada has recommended to PHAC that an annual administrative census of child welfare agencies be undertaken. Because each province and territory in Canada is responsible for its own child welfare legislation, the data received from each jurisdiction is expected to vary widely in both structure and content, which presents a significant challenge in terms of coding the data once they are received. There are difficulties with human coding of child maltreatment data that we believe can be alleviated by incorporating machine learning techniques into the coding process. Once we begin receiving data, we will develop, test, and maintain a variety of machine learning algorithms in order to create effective auto-coders that can be used to reduce the need for manual coding.

Acknowledgements

We would like to thank Lil Tonmyr (PHAC) and Wendy Hovdestad (PHAC) for their insight and their valuable feedback on this project.

References

Danso, S., E. Atwell, and O. Johnson (2014), “A comparative study of machine learning methods for verbal autopsy text classification”, preprint, (retrieved from arxiv.org/abs/1402.4380).

Marucci-Wellman, H. R., H. L. Corns, and M. R. Lehto (2017), “Classifying injury narratives of large administrative databases for surveillance—A practical approach combining machine learning ensembles and human review”, *Accident Analysis and Prevention*, pp. 359-371.

Measure, A. (2014), “Automated coding of worker injury narratives”, *JSM Proceedings, Government Statistics Section, American Statistical Association,* pp. 2124-2133.

Potter, D., W. Hovdestad, and L. Tonmyr (2013), “Sources of child maltreatment information in Canada”, *Minerva Pediatr*, 65, pp. 37-49.

Public Health Agency of Canada (2010), *Canadian Incidence Study of Reported Child Abuse and Neglect 2008: Major findings.*

Tonmyr, L., A. Asokumar, W. E. Hovdestad, M. Shields, J. Laurin, and L. Burnside (2018), “Can coders abstract child maltreatment variables from child welfare administrative data and case narratives for public health surveillance in Canada?” paper presented at the International Society for the Prevention of Child Abuse and Neglect, Prague, Czech Republic.

Zhang, Y., I. Marshall, and B. C. Wallace (2016), “Rationale-augmented convolutional neural networks for text classification”, *Proceedings of the Conference on Emperical Methods in Natural Language Processing*, pp. 795-804.