Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2006 : Methodological Issues in Measuring Population Health

2006

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2006
Methodological Issues in Measuring Population Health

Application of Statistical Disclosure Methods to the Canadian Hospitals Injury Reporting and Prevention Program Database

Ann Brown and Margaret Herbert[1]

Abstract

We describe statistical disclosure control methods (SDC) developed for a public release Canadian Hospitals Injury Reporting and Prevention Program (CHIRPP) micro-data file. CHIRPP is a national injury surveillance database managed by the Public Health Agency of Canada (PHAC). After describing CHIRPP, the paper includes a brief overview of basic SDC concepts, as an introduction to the process for selecting and developing the appropriate SDC methods for CHIRPP given its specific challenges and requirements. We then summarize some key results. The paper concludes with a discussion of the implication of this work for the health information field and closing remarks with respect to the some methodological issues for consideration.

KEY WORDS: public use micro-data file; injury surveillance; cell suppression.

1. Introduction

This paper describes the statistical disclosure control methodology developed for the public release of the Canadian Hospitals Injuries Reporting and Prevention Program (CHIRPP) micro-data file. The paper begins with brief highlights of CHIRPP and its requirements, and then includes key concepts basic in statistical disclosure control (SDC) methodology. The subsequent section presents the major steps taken to develop the CHIRPP SDC methodology. The paper includes key results for the level of suppression for the CHIRPP database and gives examples for specific fields. The paper concludes with a discussion of possible impacts of this work for the health information field and remarks on some possible methodological issues for consideration.

2. What is CHIRPP?

2.1 Canada's Injury Surveillance Program

CHIRPP is an injury surveillance program managed by the Public Health Agency of Canada (PHAC). It collects information on injuries sustained by people who report to selected emergency departments across Canada for the treatment. It is a sentinel surveillance system involving 14 hospitals (including 10 paediatric hospitals) and its purpose is to contribute to reducing the number and severity of injuries in Canada (Mackenzie and Pless). The current CHIRPP database contains information on more than 1.6 million injuries – about 100,000 events per year.

CHIRPP started in 1990 with ten paediatric hospital emergency departments (Herbert and Mackenzie). During the first decade 6 general hospitals joined the program, but 2 have subsequently dropped out. The only provinces and territories not represented are Prince Edward Island, New Brunswick, Saskatchewan and Yukon Territories.

CHIRPP operates with the support of participating hospitals. At each centre there is a CHIRPP Director – a voluntary position - usually an emergency physician who acts as a patron for the program in the hospital and community. PHAC provides funding support for a salaried CHIRPP coordinator in each centre who is responsible

[1]Ann Brown, Statistical Consultation Group, Social Survey Methods division, Statistics Canada, 15th flr., R.H.Coats Bldg, Tunney's pasture, Ottawa, Ontario, Canada K1A 0T6 (ann.brown@statcan.ca, ac.ann.brown@gmail.com); Margaret Herbert, Health Surveillance and Epidemiology Division, Public Health Agency of Canada, Jeanne Mance Building, 200 Eglantine Driveway, Ottawa, Ontario, Canada, K1A 0K9 (Margaret_Herbert@phac-aspc.gc.ca)

for distribution of CHIRPP forms to patients or parents and for collecting completed forms. The Coordinator also analyses the hospital's local CHIRPP data and provides injury data and information for the hospital and community. Some hospitals are affiliated with an injury prevention centre or a research group and CHIRPP co-Directors are sought from these groups.

2.2 CHIRPP Data

The data collected in the CHIRPP database are extracted from an open-ended questionnaire completed by patients or their parents. After data entry, each record contains about 40 variables that may be divided into the following general types. **Patient characteristics** include age, sex, postal code, occupation and industry (if the injury is work-related). **Clinical information** includes nature of injury, body part(s) involved, and disposition (whether the patient was released, admitted, etc.). This variable is used as a proxy for injury severity.

The **injury event information** collected focuses on the circumstances of the injury and is distinct from the clinical focus of injury data in other health databases. Variables include the **area** and **locations** where the injury occurred, **context** (what activity the patient was involved in), i**ntent** (unintentional, abusive or self-inflicted), **factors** that may have contributed (e.g., alcohol or drugs, products that were misused/ malfunctioned, sports, adverse environmental conditions, etc.), the **type of energy transfer** involved (mechanical, electrical, chemical, etc.). There is also a full text narrative about the injury event.

2.3 CHIRPP Data Collection and Capture

Data Acquisition for CHIRPP follows the following steps:

- Patient or parent completes questionnaire. Alternately administrative or triage staff completes form. In some centres coordinators extract data from the chart when patients are unable to complete the form;
- Forms sent to PHAC;
- Coding and electronic data entry are done at PHAC by a staff of highly trained coders who translate open-ended information into numerically coded variable and compose the narrative; and
- Basic logic / data quality checks are done at data entry - such as disallowing dates of injury that happen before the date of birth.

The National database is a relational database residing on an Oracle platform which is housed on a PHAC server. After the completion of the above steps, participating CHIRPP hospitals are sent regular updates of their own data as ACCESS datasets.

2.4 CHIRPP Data Uses and Users

CHIRPP data are used in a variety of ways. The detail and information on the injury event circumstances complements information from clinically oriented databases such as mortality or hospitalization. It provides information on types of injuries not defined by the standard health codes in the International Classification of Diseases (e.g., injuries associated with specific sports or those associated with specific products such as power tools).

The CHIRPP data provide detailed information on injury events to inform preventive efforts. For example, PHAC uses the data as a basis for injury reports, publications and for information posted on its website. PHAC responds to queries from researchers and other stakeholders, and CHIRPP information is also used on its interactive web query application - Injury Surveillance On-Line (ISOL).

Users of CHIRPP data include the Federal Government, non-government organizations involved in safety and injury prevention, public health units, health professionals, researchers, students, the private sector, media and the public.

2.5 CHIRPP Public Use Micro-data File

CHIRPP developed a Public Use Micro-data file (PUMF) of its database in response to requests for increased and broader access to its data. The PUMF will be a more useful source file for ISOL, one that uses emerging methods and technologies. It will provide data access to a wider range of users – one that will allow access without meeting the strict criteria now imposed on researchers who currently apply for data access.

There are the two fundamental values that must be addressed in developing the PUMF - the protection of the privacy of the individuals whose injuries are detailed in the database, and ensuring the provision of useful and relevant data. In order to generate the CHIRPP PUMF we required that the methods be:

- Systematic,
- Consistent with best practices in disclosure control,
- Resource efficient in terms of staff time, and consultation fees,
- Replicable for each new year of data, and
- Easily modifiable to accommodate future changes in the CHIRPP program and database.

In order to meet the specific requirements of certain users, PHAC imposed restrictions that limited the choice of disclosure control methods used to create CHRIPP PUMF. A decision was made that data from as many records as possible be used and that there be no data modifications within records, thereby excluding any data perturbation methods.

3. CHIRPP Statistical Disclosure Control (SDC) Methodology

3.1 Types of Data and SDC Methods

The choice of appropriate SDC Methods depends on the type of data to which the methods will be applied. The following indicates the types of data and describes the type of data that is contained in CHIRPP:

- **Qualitative and Quantitative Type of Data**. The CHIRPP data are qualitative, that is, each field is categorical with a finite number of categories. Therefore methods, such as rounding, that are appropriate to quantitative (numerical) data, such as income for which there are a large number of categories, were not needed.
- **Tabular data**: Methods appropriate to summarized counts presented in tables were not needed as the CHIRPP files are all individual records of injury events or micro-data.
- For **micro-data** two main groups of SDC methods are available.

    - Generally, **data perturbation methods** involve modifying data with the goal of preserving summary statistical properties such as means and variance while retaining the individual data as much as possible. These methods include randomly adding noise or randomly perturbing, and data swapping (with care given to limit bias) “synthetic” micro-data. However, these methods were not used due to the decisions noted earlier that were made that no data modification was to be made to the CHIRPP data file.
    - Thus only **Data reduction or restriction** methods were available. These are basic disclosure control methods consisting of coarsening or reducing the information. While reducing the available information limits the usefulness of the data, the data provided are as reliable and accurate as the original data. In general, these methods include:
      . Removing records for very small sub-populations or critical highly identifiable records,
      . Removing identifying fields such as names and addresses,
      . Reducing the level of detail. This involves grouping categories, recoding of some variables, sampling or sub-sampling, treatment of outliers/rare categories, top and bottom coding,
      . Suppressing data values on specific records.

3.2 SDC Concepts

The following is a brief overview of some basic and key concepts for SDC methods. These are from the (external version) of Statistics Canada's (STC) Handbook for Creating Public Use Micro-Data files (2006). This document is based on STC's many years of experience releasing STC's Social and Household Survey PUMF's, striking a balance between the goals of maintaining the quality and usefulness while protecting confidentiality, as necessary by the STC Act.:

- **Re-identification** or **disclosure of an individual** may occur when a publicly released micro-data file either leads to (or allows) the identification of an individual. This is may occur even when direct identifiers such as name, telephone number, or other identification numbers (that could be used to link to other files e.g., Social Insurance Numbers) have been removed from the file. No record on the file should lead to the identification of an individual.
- **Indirect identifiers** are those fields contained in the database such as geography and (demographic) characteristics of an individual. The database may also contain **sensitive** fields containing personal information such as health, crime or income. So, key variables, a combination of which could allow for re-identification, include indirect and sensitive variables.
- **Key Variables** include those fields/variables on the released micro-data file that may be used for re-identification purposes. For example these include indirect identifiers and other fields relating to an individual that when used in combination with the indirect identifiers may lead to re-identification. These other fields are the sensitive fields and are specific to the database – the CHIRPP sensitive fields will be listed shortly in the tables of results.
- **Disclosure risk** methods measure the risk of disclosure contained within a database. They typically use the set of key variables and involve the calculation of **unique**, doubles, triples and quadruples as a portion of the entire database, where a unique record is a record that occurs only once for the specific value of the key variables under consideration.
- **Suppression** or treatment of risk involves specific values of cell combinations of key fields. That is, it involves deciding on a specific number of records or **threshold** with the same values or the combinations of the key variables that pose an unacceptable risk of re-identification. The threshold for the treatment risk can be set at 3 or higher. The suppression treatment is applied to the values of the key variables of the records not meeting the threshold.

3.3 Notable Characteristics of CHIRPP

In general, the features and unique challenges posed by the CHIRPP file included the amount, type, level of detail, as well as the indirect and sensitive variables it contained. The specifics of these are described in the next section. However, the most striking characteristic was its skewness and the presence of some rare records. This arises as the CHIRPP contains information on injuries at participating hospitals, which are not in all provinces/territories. That is, the participating hospitals are not randomly chosen representative sample of hospitals. As a consequence the following groups are under-represented: older teenagers and adults who seek treatment in general hospitals; rural populations (most CHIRPP hospitals are located in large urban centres); and aboriginal populations living in rural and remote areas. As an example, Age, Industry and Occupation distributions are not the same as the general population. Only 20% of CHIRPP records are for persons aged 15 and older, so variables such as industry occur as a small proportion of the already small subset of persons aged 15 or older.

3.4 CHIRPP SDC Methodology

We describe the process and the main steps followed to select the appropriate methods for CHIRPP, given its characteristics and requirements, and how these were applied and optimized. These methods were applied to each year of the CHIRPP data.

The initial steps are the examination and analysis of the CHIRPP data to identify small high risk sub-populations, high risk fields, and any group of key and sensitive variables that might pose a risk for re-identification when used alone or in combination. That is, the content of the CHIRPP micro-data file was analyzed according to:

- **CHIRPP geography:** postal code and hospital code that identifies the province/territory and municipality of the hospital.
- **Direct identifiers** clearly identifying the individual having the injury event: CHIRPP contains a field with the first 3 characters of last name of the person experiencing the injury.
- **Indirect identifiers** include fields that characterize an individual. For CHIRPP, these are age, gender, occupation, industry, and language spoken.
- **Sensitive information**, such as injury event descriptions of medical and other circumstances, due to its nature and/or rare occurrence.
- **Level of Detail**. The univariate frequency counts of the categories contained in each CHIRPP field were produced.

Subsequently, data reduction methods appropriate for CHIRPP were developed. These involved the:

- Exemption of Rare records
  Fatality records were exempted from the file, given their very small numbers—roughly 50 per year. Given that they compose a very small subpopulation within CHIRPP, there is the unacceptable high risk of unintentional disclosure. Fatalities are often well known due to media publicity and in the community surrounding the event and due to media publicity — e.g. major rail crashes, motor vehicle collisions, etc. Furthermore, information about fatalities is readily available from other sources including obituaries, public summary statistics, or on-line death databases. Combining the publicly available information with the information available on CHIRPP may re-identify an individual and reveal new personal information about the individual or the injury event. For example, deaths involving trains are rare (often less than 10 per year in Canada) and individuals may be readily re-identified by age and sex. A CHIRPP record involving a train fatality might also reveal additional information such as suicidal intent, involvement of alcohol, or harmful peer interaction.

- Exemption of Fields
  Fifteen fields were exempted from the database due to their unacceptably high risk, which include:
  - Direct identifiers: partial name, hospital code, chart number
  - Geographic fields: postal code, hospital code
  - Dates: birth, injury event, ER visit
  - Rare Medical circumstances: intent
  - Narrative text fields: Injury event description due to the impracticality of manually reviewing the large volume.

- Release of Key Fields
  Key fields were able to be released through the use the following disclosure control methods:
  - **Coarsening**: regrouping of categories with small counts. For example, these included regrouping of rare injuries such as amputations. Subject matter expertise was essential to maintain usable and meaningful categories. Before and after frequency counts were done as a check.
  - **Standard categories**: "Other" and "Not Stated" are valid original code values for all variables. Information concerning some variables in the CHIRPP is not applicable to all injuries and in such cases is deemed "not stated", e.g., vehicle seating position for injury events not involving vehicles. The "Other" category is used for circumstances not covered by existing code categories.
  - **Cell exemptions:** Combinations of the key fields were examined and treated to suppress cell values for rare combinations using minimum counts per cell. This includes records in rare cells for the five-way combination of the indirect demographic identifiers, four-way combinations involving age and gender and all combinations of 7 highly sensitive fields (body part, breakdown event, context, location, mechanism factor, mechanism of injury, and nature of injury); 4-way combinations for age and gender and all combinations of 4 less sensitive fields (disposition, injury group, safety device 1, and vehicle seating position). All lower dimension combinations of each of these were also treated. Monitoring of the record level cell exemptions was done by defining a flag to indicate if at least one of the fields was suppressed.

The last steps involved testing the pilot and optimizing the approach. Further coarsening of the detail of some key variables was done after the cell suppression routines developed for the pilot revealed an unacceptably high level of suppression for some useful circumstance variables. In the pilot, many combinations of the key variables in the cell suppression routines were ad hoc and did not include any demographic fields. Thus, optimization involved, first, systematically specifying the cell suppression routines to include all possible combinations of the key variables according to their sensitivity, and secondly, fine-tuning their order more systematically within the cell suppression routines. Optimization resulted in a reduction of the number of combinations of the key variables examined in the cell suppression routines as well as increasing the amount of released data.

3.5 Key Results

Table 1 shows the reduction in detail of the number of categories for the CHIRPP key fields. It compares the number of categories in the original file before and after coarsening. Mechanical Factor, the most detailed CHIRPP key field, was reduced to 13 categories from the original 629 possible categories.

TABLE 1: Reduction in Categories	# OF POSSIBLE VALUES	
KEY VARIABLE	BEFORE	AFTER
Age group	7	5
Body Part	31	15
Breakdown Event	29	10
Context	49	10
Disposition	10	5
Home Language	31	4
Gender	3	3
Industry	36	6
Injury Group	7	4
Location	57	7
Mechanism	33	9
Mechanical factor	629	13
Nature of injury	38	11
Occupation	187	4
Safety device	10	6
Vehicle Seating position	22	4

Table 2 compares the results of the applying the cell exemption step of the CHIRPP SDC methodology to the original un-coarsened CHIRPP data and the optimized approach for the 2002 CHIRPP year. It is to be noted that the level of suppression was much higher for the un-coarsened data. That is, the data quality, as measured by the amount of data than can be released, using the optimized approach was much higher when the coarsened data were used in the optimized approach. For example, the additional suppression for the Context field was 23.9% when un-coarsened data were used and 1.8% when coarsened data were used in the optimized approach. Further, for the 2002 dataset the overall level of suppression (the number of records with at least one suppressed field) was reduced from 72% to 7%.

TABLE 2: % INC. NOT STATED	% RECORDS	
KEY VARIABLE	2002 orig., no coarsening	2002 Optimized
Age group	22.4	7.0
Body Part	37.5	2.6
Breakdown Event	45.8	1.9
Context	23.9	1.8
Disposition	5.1	0.1
Home Language	5.2	5.4
Gender	5.7	0.2
Industry	1.0	0.2
Injury Group	1.2	0.1
Location	44.4	1.1
Mechanism	13.8	1.5
Mechanical factor	62.6	2.3
Nature of injury	32.7	1.4
Occupation	1.7	0.2
Safety device	2.5	0.1
Vehicle Seating position	1.8	0.1
OVERALL -RECORDS	72.7	7.0

Tables 3A and 3B show results of applying the cell exemption step of the CHIRPP SDC methodology for the distribution of the categories for two CHIRPP fields, as an illustration of the information loss resulting from the suppression treatment. While the initial non-response rate was very low for 'Nature of Injury', and much higher for 'Occupation', the increase in 'Not Stated' was marginal, with the all of the categories having similar distribution before and after cell suppression.

TABLE 3A: IMPACT ON CELL VALUES

CHIRPP FIELD: "OCCUPATION"	BEFORE %	AFTER %
Managers/Professionals/Para Professionals	0.6%	0.5%
Trades/Clerks/Sales/Machine Operators/Drivers	1.3%	1.3%
Labour/Construction Workers	0.9%	0.8%
Not Stated	97.2%	97.4%
TOTAL	100.0%	100.0%

TABLE 3B: IMPACTON CELL VALUES

CHIRPP FIELD: "NATURE OF INJURY"	BEFORE %	AFTER %
Superficial	10.3%	10.2%
Open Wound	17.8%	17.7%
Sprain or Strain	9.9%	9.8%
Eye Injury	2.4%	2.2%
Minor Head Injury/Concussion/Intracranial	9.0%	8.9%
Foreign Body	2.8%	2.6%
Thermal/Electrical/Poisoning/Toxic Effect	3.5%	3.3%
Soft Tissue Injury	12.2%	12.0%
Fracture/Dislocation/Pulled Elbow	24.4%	24.3%
Other	7.6%	7.6%
Not Stated	0.0%	1.4%
TOTAL	100.0%	100.0%

4. Discussion

We believe that this is the first on the databases managed by PHAC for which a PUMF has been prepared. It's been a learning process for us and much that we have learned is applicable to other health data holdings. Emerging best practices in disclosure control are already helping PHAC develop guidelines for release of information from health data sets. Many of the challenges encountered in producing the CHIRPP PUMF could have been avoided or minimized if the creation of a PUMF version had been foreseen when the database was designed. As use of PUMFs becomes more widespread, database design and development will need to consider the benefits of trying to minimize collection of personal data and infrequently used variables and by minimizing the number of categories for specific variables.

The application of SDC methodologies to CHIRPP was not a straightforward process. It involved a series of trials and adjustments aimed at achieving a high level of data released while maintaining a high standard for privacy protection. As an example the threshold level for univariate frequencies initially used to group the detailed categories into summarized categories was 0.5% of the file. Testing at this level produced a file with unacceptably high levels of suppression. Due to the skewness of the CHIRPP data, this threshold for the coarsening the detailed categories of the CHIRPP fields was finally established to be about 2.5%.

It is important to recognize that application of disclosure control procedures to CHIRPP or any other database minimizes the risk of re-identification of individuals; it does not eliminate the risk. This is well illustrated by an example - a hypothetical CHIRPP case: A 9 year old girl, who is receiving money to deliver flyers, is attacked by a dog, as she is walking along the sidewalk, bitten in the face and admitted to hospital. The corresponding CHIRPP record (after roll up of categories) would contain the following information items all under different variables: 5-9 year old/ female/ injured while working for pay/ a cut or puncture or bite/ involving an animal/ happened on a sidewalk or parking area or driveway/ patient admitted to hospital.

All of these information items are commonly reported in CHIRPP, but the combination occurs rarely and each additional piece of information makes it easier to link the CHIRPP record with what is known or reported about the injury event. If this incident was reported in the media there is a residual risk that a link could be made to the person. The risk would be much greater if this injury was fatal - there are only a few child deaths involving dog bites each year and they are widely reported. For this reason fatal injuries were suppressed as part of the disclosure control on CHIRPP.

5. Concluding Remarks

The CHIRPP SDC methodology assessed and treated risk within the CHIRPP file in a comprehensive, systematic, and efficient manner, consistent with accepted SDC best practices. These SDC methods should not be applied indiscriminately. They need to be applied on a case-by-case basis because the science of SDC Methods is evolving and the process for applying the chosen appropriate methods is an art that involves making the best possible decisions with the tools and information that is available at that time.

With increasing demand from users, advances in technology, and the science of SDC methods work is needed to monitor and evaluate the application, and results of using SDC methods. This is especially the case in determining the level of acceptable risk that may still be present due to factors that are external to the database. Ultimately the acceptable risk level is a decision to be made by the owners/managers of the database after due consideration of expert advice.

References

Boudreau, J. R. (2005), "Data Swapping is not the Panacea", *Proceedings of Statistics Canada Symposium 2005: Methodological Challenges for Future Information Needs*, Section 1.

Herbert, M. and Mackenzie, S. G. (2004), "Injury Surveillance in Paediatric Hospitals: the Canadian Experience", Paediatrics and Child Health, 2004, Volume 9, Number 5, pp. 306-308.

De Waal, A. G. and Willenborg L.C.R.J (1998), "Optimal Local Suppression in Micro-Data", *Journal of Official Statistics Sweden*, 14, pp. 421-435.

Mackenzie, S. G. and Pless, I. B. (1999), "CHIRPP. Canada's Principle Injury Surveillance Program", *Injury Prevention*, 1999, Volume 5, pp.208-213.

Santos, M. J. (2005), "Statistical Disclosure Control: Legal Framework and Methodological Aspects ", *Proceedings of Statistics Canada Symposium 2005: Methodological Challenges for Future Information Needs*, Sec. 16

Skinner, C. J. and Holmes, D.J. (1998), "Estimating the Risk of Re-Identification Risk per Record", *Journal of Official Statistics Sweden*, 14, pp. 361-372.

Statistics Canada (2003), catalogue no. 12-587-XPE, "Confidentiality and Disclosure Control" in *Survey Methods and Practices*, Chapter 12 section 5, pp 269-278

Statistics Canada (June 2006), "Handbook for Creating Public Use Micro-data files",.

U.S. Office of Management and Budget, Federal Committee on Statistical Methodology, (December 2005), "Chapter II – Statistical Disclosure Limitation Methods: A Primer, Section F. Microdata "*Report on Statistical Disclosure Limitation Methodology*, pp 25-33

Acknowledgement

We would like to acknowledge the valuable contributions of Bev Cleary from the Public Health Agency of Canada and of Lori Stratychuk from Statistics Canada.