Use of Probabilistic Record Linkage for the Canadian Cancer Registry

Jocelyne Marion and Brad Thomas, Statistics Canada
R.H. Coats Building, 16th floor, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6

Abstract

Cancer is a major cause of deaths in Canada. Both its detection and treatment place great burdens on the Canadian Health Care system. It is vital for the Canadian government to be able to estimate accurately the incidence of cancer in Canada each year, as well as be able to detect trends which indicate increases or decreases, in order that sufficient health care funds be made available. Cases of cancer incidence in Canada are reported to the Provincial/Territorial Cancer Registries (PTCRs), who in turn pass this information onto Statistics Canada, for inclusion onto the Canadian Cancer Registry (CCR), a patient-oriented database that has cancer incidence in Canada recorded since 1992.

This paper focuses on estimating the incidence of cancer in Canada and the use of probabilistic record linkage to reduce cases of both under-estimation and over-estimation of cancer incidence. There are two linkages done in annual CCR production: i) an internal linkage of the CCR to itself to detect possible duplicates of patients or tumours; and ii) an external linkage of the CCR to the Canadian Mortality DataBase (CMDB) to find cases of cancer that were not detected until the time of death

Statistics Canada uses its internally-developed Generalized Record Linkage System (GRLS) in order to carry out probabilistic record linkage. Version 1 of this system, developed back in the 1970s resides on a mainframe computer, and is the version still used for production in the CCR. The current version of this system, Version 4.4, is UNIX-based. Linkage production for the CCR is being converted from V1 to V4.4, which also provides an opportunity to evaluate how best to improve the linkage work being done.

Key Words: probabilistic record linkage, internal linkage (unduplication), external linkage

1. Introduction:

This paper will begin with a description of the cancer situation in Canada. Subsequent sections will then describe how national-level counts of cancer cases are compiled, and then after some background information on probabilistic record linkage in general, how probabilistic record linkage is used for the cancer program at Statistics Canada.

2. Cancer in Canada

Cancer is a major cause of deaths in Canada. Cancer and heart disease together represent the two leading causes of deaths among Canadians. Statistics Canada is responsible for the collection of cancer incidence and deaths in Canada so that policy makers may have access to the most accurate and up-to-date counts possible of the prevalence of cancer in Canada.

Statistics Canada collects malignant and invasive cancers in its count, but publishes malignant cancers only. These are cancers which, if left untreated, could cause death. From the 10th edition of the International Classification of Diseases (ICD), this would be cancers in the range from C00 through C80 inclusive. The only exception is non-melanoma skin cancer (basal and squamous), since only one of Canada's 10 provinces reports it.

In addition, further to the ICD system of coding diseases, changes to the coding system can also change the scope of what is counted and what is not. Consider the ICD-O system, which is the coding system for oncology, where oncology refers to further specification of cancer for research purposes. The latest edition is the ICD-O-3, in which borderline ovarian cancer (which had been included in ICD-O-2), is no longer considered to be malignant, so as a result Statistics Canada no longer counts borderline ovarian cancer. So as a result, in the move from ICD-O-2 to ICD-O-3, it will appear as though there has been a drop in ovarian cancer in Canada, whereas in reality, the borderline ovarian cancer has merely been dropped as being out-of-scope.

Here are some highlights from the projected picture of cancer in Canada for 2004. It has been estimated that there will be 145,500 new

malignant cases of cancer, and 68,300 deaths due to cancer. These estimates represent incidence and mortality rates of 0.46% and 0.21%, respectively, based on a mid-2004 population projection of 31,917,199 (medium scenario, provided by Demography Division of Statistics Canada). The most frequently diagnosed cancers will be prostate for men, and breast for women. The leading deaths due to cancer will be lung, followed by colorectal, for both sexes.

Canada ranks among the nations of the world with the highest rates of lung and colorectal cancer, as well as prostate cancer among men, and breast cancer among women. At the other end of the spectrum, Canada ranks among the nations with lowest rates of stomach, liver and cervical cancers as well as cancer of the esophagus.

3. Counting Cancer in Canada

Counts of cancer incidence (i.e. new cases of malignant cancers) are kept by the various Provincial/Territorial Cancer Registries (PTCRs) across the country. There is one PTCR in each of Canada's 10 provinces, and also one in each of the three northern territories. Not all PTCRs were established at the same time. The earliest of these PTCRs date from the 1930s from the provinces of British Columbia and Saskatchewan. Part of the reason why the PTCRs came into existence at different times across the country is that health care is a provincial/territorial responsibility, not a national one. Furthermore, the reporting of cancer incidence is not required in all provinces/territories, although most of them do have a legal requirement for it.

Seven provinces (Newfoundland, Nova Scotia, Ontario, Manitoba, Saskatchewan, Alberta, British Columbia) have cancer agencies with a legislated responsibility for cancer control. In provinces and territories without cancer agencies, cancer control is the responsibility of the Departments of Health. In Quebec, there is a cancer control program guided by an advisory council. In New Brunswick, cancer care is managed through eight hospital corporations and two cancer centres. In Prince Edward Island, the Department of Health and Welfare is responsible for cancer control. [1]

With the increasing use of computer technology by the 1960s, it appeared to be feasible to have a national system to compile these counts. In 1969, Statistics Canada (then known as "The Dominion Bureau of Statistics") launched the National Cancer Incidence Reporting System (NCIRS). This was an incident-oriented system used up to 1991. By 1992, this was replaced with the Canadian Cancer Registry (CCR), a person-oriented system which Statistics Canada still uses.

The PTCRs obtain their counts from a variety of different sources:

- cancer clinic files
- radiotherapy reports
- in-patient hospitals
- out-patient clinics
- private hospitals
- pathology and other lab reports
- autopsy reports
- radiology and screening program reports
- private-practice physicians
- reports on deaths from cancer from Registrars of Vital Statistics

With this many sources, it is not surprising that some duplication manages to escape detection by the PTCRs. Statistics Canada has a role to play in that, but more on that later.

Figure 1, which appears after the references, (adapted from Thomas, 1999) shows an overview of how the PTCRs submit their data to Statistics Canada, and how Statistics Canada communicates with the PTCRs concerning cases which fail the extensive edits to which the data from the PTCRs are subjected. This shows a major advantage over the typical scenario with administrative data, in that the recipient of the administrative data can send data back to the source for correction. This partnership between Statistics Canada and the PTCRs is part of the shared goal of better data quality.

[1] Detailed information regarding the legal context for cancer surveillance including a review of relevant legislation, codes, policies and procedures and common law in Canadian jurisdictions is available in *Use of Cancer Patient Information for Surveillance Purposes* prepared by the Health Law Institute at the University of Alberta and the Centre de recherche en droit at the Université de Montréal.

Better data quality is also the goal of the Data Quality Management Committee. Statistics Canada co-chairs this committee, which oversees data quality for the CCR. It reports to the Canadian Council of Cancer Registries. Members include not only representatives from the PTCRs, but also representatives from Statistics Canada and Health Canada as well as a pathologist. This committee meets bi-annually, hosted by Statistics Canada, in order to discuss issues of common concern, as well to find ways to standardize data capture and coding practices across Canada so that data can be collected and disseminated more quickly.

Timeliness is a big issue with data for the CCR. The process of collecting data from the PTCRs, and then compiling them at the national level after they have been through the exhaustive edit and feedback process results in a typically long delay. For example, the publication *Canadian Cancer Statistics 2004* uses CCR data from 2001, since that was the most recent year for which incidence and mortality data were available on a national level. The process is further delayed by the annual CCR record linkages that are done in order to control both over- and under-counting.

3.1 CCR Internal Linkage

As mentioned earlier, given that the PTCRs collect incidence data from such a wide variety of sources, duplication can occur whereby the PTCR assigns a different identification number to two related records as if they belonged to different people. In addition, there is duplication possible between the PTCRs, because people migrate from one province to another, and it is possible that such a person could have a cancer incidence in both their current and former provinces.

The goal of the CCR Internal Linkage is to link the CCR to itself in order to detect duplication both within the PTCR and between the PTCRs. Once the linkage is done, reports on duplicates detected are then sent to the PTCRs for verification and resolution. This Internal Linkage seeks to prevent over-counting of cancer incidence by ensuring that each cancer patient is represented just once.

3.2 CCR External Linkage (or Death Clearance)

Once the CCR has been unduplicated, the next step is to link the CCR to the Canadian Mortality Data Base in order to verify death information on the CCR and to find cases of cancer which are detected only at the time of autopsy, and thus which may have escaped detection by the PTCRs. This linkage, known as the External Linkage, or Death Clearance, seeks to prevent under-counting of cancer incidence.

4. Probabilistic Record Linkage

Both the Internal Record Linkage and Death Clearance are accomplished with probabilistic record linkage. In a typical record linkage procedure involving more than one file, common variables between the files are identified to be used as linkage variables, and then the values of these linkage variables are compared for a pair of records (each one from a different file) in order to make a decision as to whether or not the records refer to the same entity (e.g. person).

In probabilistic record linkage, the record pair does not have to agree on all values of the linkage variables in order to accepted as a true link. Instead, a weight, or score, is assigned to the values on which the linkage variables do agree (either fully or even partially, if the probabilistic record linkage system is sufficiently sophisticated). These weights, or scores, are added up over all the linkage variables to obtain a total weight, under the assumption of statistical independence between the comparison rules (Fellegi and Sunter, 1969). If this total weight exceeds a certain threshold, then the record pair is kept as a true match.

Real probabilistic record linkage applications will typically include other features such as negative weights for disagreements, weights of any sign (positive, negative, or zero) for missing value comparisons, and blocking. Further details are in Statistics Canada (2001), but blocking deserves some further explanation here.

Blocking, also known as pockets or initial criteria, involves reducing the number of potential pairs to have to evaluate. For example, if we have File A with N_A = 10,000 records, and Field B with N_B = 100,000 records, then we would have $N_A * N_B$ = 1,000,000 record pairs to have to evaluate, which could prove excessive for our computer system. But if we break up File A into *n* non-overlapping components of

size A_1, A_2, …, A_n so that $N_A = \Sigma_k A_k$ (e.g. A_1 is all the records from File A where patient surname starts with an "A"), and similarly break up File B into *n* non-overlapping components of size B_1, B_2, …, B_n, so that $N_B = \Sigma_k B_k$, using the same criteria used for File A, then we can restrict record pairs to those which come from the same subset of *A* and *B*, e.g. A_k and B_k. So then we have just $\Sigma_k A_k * B_k$ comparisons to make, which is considerably less than $(\Sigma_k A_k) * (\Sigma_k B_k) = N_A * N_B$.

Probabilistic Record Linkage systems will also allow for frequency weighting as opposed to global weighting. In global weighting, one may assign for example a weight of 100 for an agreement on patient surname, regardless of value. In frequency weighting, one can vary the weight so that a more common name, such as "Smith" gets a lower value than a more rare name such as "Stankewicz".

Statistics Canada has developed its own internal probabilistic record linkage system, currently known as GRLS for Generalized Record Linkage System (formerly known as GIRLS, or CANLINK). The first version of this system appeared in the 1970s and was mainframe based. It is still used, although unsupported. It offered character variable comparison on both whole strings and sub-strings, plus a nickname dictionary (e.g. to relate "Bob" to "Robert"). Numeric variable comparisons allowed for comparisons on exact values and tolerance factors. Users were also free to add their own user-defined rules in PL/1 code in case these built-in comparison functions were not versatile enough to satisfy their needs.

The most current version of the system is v4, which is UNIX-based and uses the Oracle relational database management. In addition to the character variable comparison functions found in v1, v4 also offers encoding schemes such as the New York State Information Intelligence System (NYSIIS), whereby names such as "Smith" and "Smythe" can be linked up because of their common NYSIIS value of "SNAT". A string comparator function, which offers a numerical measure of how "close" two character strings are, is also available, which follows the methodology of Winkler (1990). User-defined rules can be supplied again, but this time in C code. In addition, by virtue of using Oracle, variables in date format can be used in special date comparison rules.

5. GRLS and the CCR

Currently, both Internal Linkage and Death Clearance are done in production using v1 of GRLS. The challenge before us is to convert the v1 application to v4, particularly because v1 is mainframe-based, and Statistics Canada is exploring a contingency plan to eliminate mainframe usage in the next 5 years.

5.1 Internal Linkage

For the linkage of the CCR against itself, the v1 application has 28 linkage rules, of which 17 are user-defined PL/1 code. A conversion to v4 raises a number of issues, such as data storage capacity, format conversion (Oracle tables for v4 versus flat files for v1), processing data volume on a smaller platform (UNIX server for v4 versus the superior number-crunching power of the mainframe for v1). So as a first step, it was decided to try just a translation from v1 to v4 in order to assess the impact of a move from mainframe to UNIX server.

5.1.1 Internal Linkage – the Translation

There are a number of other GRLS v1 applications at Statistics Canada, and under a contingency plan of removing mainframe, they would all have to be converted to v4. A strict translation from v1 to v4 would provide us with a benchmark of the impact of the move from mainframe to UNIX server for the CCR application. This would still provide valuable information for other GRLS v1 applications which need to be moved to v4.

For the CCR Internal Linkage, the most resource-intense part was the translation of the 17 user-defined rules from PL/1 code to C code for running on GRLS v4. Once translated, the rules also had to be tested for accuracy. With the translation completed, we proceeded with a test run for a subset of Ontario records.

We took a set of 1,381 records, for which GRLS v1 made linked groups with 1,050 records. With our v4 translation, using the same input, we constructed groups using 1,044 records (99.4%), which is a very good reproduction rate. Furthermore, of the 1,044 records used, 955 had the same total global weight in both the v1 and v4 runs. Of those that differed, most differed by

only small amounts. This was very encouraging, despite the fact that we limited our test just to global weights. (There were further conversion challenges that awaited us if we chose to try to reproduce frequency weight results.) At this point, given the amount of resources that had already been expended to get to this point, we did not pursue any tests with bigger data sets or with frequency weights, but instead moved on to the complete re-write.

5.1.2 Internal Linkage – the Re-write

For the re-write in v4, we take advantage of all the advanced features (the "bells and whistles") that v4 offers over v1. Since the translation of user-defined rules took the most resources in the previous phase, it was hoped that a complete re-write would have less reliance on user-defined rules in order to achieve the same results that v1 gave. We were successful in this goal: the complete re-write used 25 rules, and just 3 of these still had to resort to user-defined code.

We then did a further test with a set of Ontario records. Running all records from Ontario through Internal Linkage in current production involves about 450,000 records for the whole 1992-2002 period. We took a set of 1,003 records that were identified from production as being duplicates. That means that from the initial set of duplicates found from v1, these were the final set of duplicates identified after the PTCRs had given us their feedback. We put these through the v4 re-write, and managed to achieve results that were closer to the desired end result than the original v1 run did. Encouraged by this success, we are next moving on to a bigger test. But this provides us another problem – access to larger data storage/processing capacity.

Although one of the goals of this v1 to v4 conversion exercise is to wean CCR production off of the mainframe, we had actually used UNIX on the mainframe in order to conduct all the tests which we had done to date, due to lack of access to a UNIX server which could accommodate the volume of data which we needed even just for our small tests. But even these tests proved expensive to do on the mainframe.

For the desired larger test, we have managed to locate another UNIX server which was in a down period and which therefore we could use. But already this has raised questions as to whether cut-off from the mainframe will even allow us to continue CCR production.

Space considerations aside, the main obstacle now to v4 on the server is the time needed to create the pairs. Phase 1 is the phase of the GRLS v4 run which creates pairs of records(links) under the constraints of the selected pockets and phase 2 passes the links through the rules and either rejects or keeps them. A weight is calculated for the kept links. Phase 1 takes about 1/2 hour even for as many as 98 million links (for an input of 480,000 records, after blocking is applied). Phase 2 on the other hand, easily takes an hour to process 3 million links, so it becomes important to add additional criteria to the pocket constraint in order to minimize the links produced in Phase 1. This option of adding such an additional criteria was not available in v1. So we are certainly exploring options now available in v4 to ease the transition from v1.

5.2 CCR: External Linkage

Since we are still trying to run tests which are more representative of the volume which we can expect to face in production for the re-write of Internal Linkage, we have not yet moved on to work on the CCR Death Clearance.

The current V1 application has 24 linkage rules, of which 19 are user-defined. When we do move on to the External Linkage, we will proceed directly with the re-write, since translation work has proved to be too resource-intensive. We look forward to seeing similar success with less dependency on user-defined rules in an External Linkage re-write for v4 as we found for Internal Linkage.

6. Conclusion

Although work is still continuing on the conversion for the CCR Internal Linkage, it is safe to say that the v4 application will be preferred to the v1 application, and this will likely also be the outcome for the CCR External Linkage, or Death Clearance, as well. The question remains simply how best to set up the system to run the v4, given the UNIX and Oracle requirements.

7. References:

Fellegi, Ivan P. and Sunter, Alan B. (1969). "A Theory for Record Linkage". *Journal of the American Statistical Association*, Vol 64, pp 1183 - 1210.

Gaudette, L. A.; Labillois, T.; Gao, R.-N.; Whittaker, H.. (1996) "Quality Assurance for the Canadian Cancer Registry", Proceedings from the 1996 Statistics Canada Symposium, *Non-sampling Errors*

National Cancer Institute of Canada. Canadian Cancer Statistics, 2004, Toronto, Canada, 2004

Statistics Canada (2001). *GRLS Concepts Guide*. Systems Development Division, Ottawa, Canada.

Thomas, B. (1999). Periodic Review of the Cancer Program: Canadian Cancer Registry and Canadian Cancer Data Base: Methodology Review, Statistics Canada Internal Report,

Winkler, W. E. (1990). "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp 354-359.

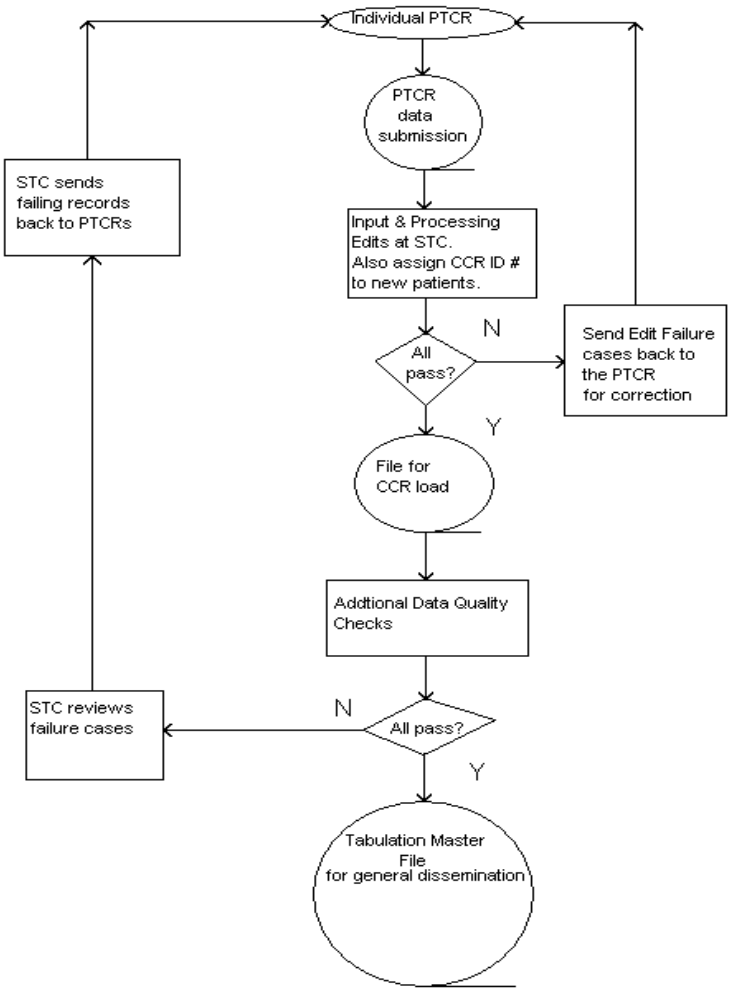

Figure 1. Flow diagram for data submission by PTCRs to Statistics Canada (STC)