

## A Comparison of Direct Match and Probabilistic Linkage in the Death Clearance of the Canadian Cancer Registry

*Tony LaBillois, Marek Wysocki, and Frank J. Grabowiecki,  
Statistics Canada*

---

### **Abstract**

*The Canadian Cancer Registry (CCR) is a longitudinal person-oriented database containing all the information on cancer patients and their tumours registered in Canada since 1992. The information at the national level is provided by the Provincial and Territorial Cancer Registries (PTCRs). An important aspect of the CCR is the Death Clearance Module (DCM). It is a system that is designed to use the death records from the Canadian Mortality Data Base to confirm the deaths of the CCR patients that occurred during a pre-specified period. After extensive pre-processing, the DCM uses a direct match approach to death confirm the CCR patients that had a death registration number on their record and it performs a probabilistic record linkage between the remaining CCR patients and death records. For one province, death registration numbers are not provided with the cancer patient records. All these records go directly to the probabilistic linkage. For the rest of the country, a good proportion of the cancer patients reported as dead by the PTCRs have such a number that can be used to match directly the two databases. After an overview of the CCR and its DCM, this presentation will compare the situation where the direct match is used in conjunction with the probabilistic linkage to death confirm cancer patients versus the case where the probabilistic record linkage is used alone.*

### Introduction

In combining two sources of data, it is sometimes possible to match directly the records that represent the same units if these two sources have one common unique identifier. Nevertheless, it is often not possible to find all the common units using only this approach, either because the two sources do not have a common unique identifier, or because, even when used, it is not complete for all the records on the files.

The case of the Death Clearance (DC) of the Canadian Cancer Registry (CCR) is an example of the latter. The purpose of this task is to associate cancer patient records with death certificate records to identify the individuals that are present on both files. The CCR already contains the death registration identifier for some patients, but not for all that may indeed be deceased. Consequently, the most reasonable process involves matching directly all the CCR patient records that have this information, and then using probabilistic record linkage in an attempt to couple the remaining records that could not directly match. It is our belief that this maximises the rate of association between the two files while reducing the processing cost and time. In this situation, one could also use probabilistic linkage, alone, to perform the same task. The intention of this study is to compare these two approaches.

Firstly, this paper provides an overview of the CCR with emphasis on the Death Clearance module.

Secondly, the characteristics of the populations used in the study are described. Next, the paper explains the comparisons between the two approaches (process, results and interpretations); and finally, it presents the conclusions of this study.

## Overview of the Canadian Cancer Registry

The Canadian Cancer Registry at Statistics Canada is a dynamic database of all Canadian residents diagnosed with cancer [1] from 1992 onwards. It replaced the National Cancer Incidence Reporting System (NCIRS) as Statistics Canada's vehicle for collecting information about cancer across the country. Data are fed into the CCR by the 11 Provincial and Territorial Cancer Registries (PTCRs) that are principally responsible for the degree of coverage and the quality of the data. Unlike the NCIRS that targeted and described the number of cancers diagnosed annually, the CCR is a patient-based system that records the kind and number of primary cancers diagnosed for each person over a number of years until death. Consequently, in addition to cancer incidence, information is now available about the characteristics of patients with multiple tumours, as well as about the nature and frequency of these tumours. Very importantly, since patients' records remain active on the CCR until confirmation of their death, survival rates for various forms of cancer can now be calculated.

The CCR comprises three modules: *core*, *internal linkage* and *death clearance*. The *core* module builds and maintains the registry. It accepts and validates PTCR data submissions, and subsequently posts, updates or deletes information on the CCR data base. The *internal linkage* module assures that the CCR is truly a person-based file, with only one patient record for each patient diagnosed with cancer from 1992 onwards. As a consequence, it also guarantees that there is only one tumour record for each, unique, primary tumour. The *internal linkage* identifies and eliminates any duplicate patient records that may have been loaded onto the database as a result of name changes, subsequent diagnoses, or relocations to other communities or provinces/territories. Finally, *death clearance* essentially completes the information on cancer patients by furnishing the official date and cause of their death. It involves direct matching and probabilistic linking cancer patient records to death registrations at the national level.

## The Death Clearance Module

Death clearance is conducted on the CCR in order to meet a certain number of objectives (Grabowiecki, 1997). Among them, it will :

- permit the calculation of survival rates for patients diagnosed with cancer;
- facilitate epidemiological studies using cause-of-death; and
- help file management of the CCR and PTCRs.

The death clearance module confirms the death of patients registered on the CCR by matching/linking [2] their patient records to death registrations on the Canadian Mortality Data Base (CMDB), or to official sources of mortality information other than the CMDB. These other sources include foreign death certificates and other legal documents attesting to, or declaring death (*they are added to the CMDB file before processing*).

The first major input to this module is the CCR database that is built of patient and tumour records. For every person described on the CCR, there is only one patient record, but as many tumour records as there are distinct, primary cancers diagnosed for that person. Patient records contain nominal, demographic and mortality information about the person, while tumour records principally describe the characteristics of the cancer and its diagnosis. CCR death clearance uses data from the patient record augmented with some

fields from the tumour record (the tumour record describing the patient's *most recently diagnosed tumour* when there is more than one). More details on the variables involved are available in Grabowiecki (1997) and Statistics Canada (1994).

The second main input is the Canadian Mortality Data Base. This file is created by Statistics Canada's Health Division from the annual National Vital Statistics File of Death Registrations, also produced by Statistics Canada. Rather than going directly to the Vital Statistics File, death clearance uses the CMDB as the principal information source about all deaths in Canada, because of improvements that make it a better tool for record linkage. A separate record exists on the CMDB for every unique reported surname on each Vital Statistics record -- viz.: the deceased's surname, birth/maiden name, and each component of a hyphenated surname (e.g., Gérin-Lajoie, Gérin, and Lajoie). All of the above surnames and the Surname of the Father of the Deceased have been transformed into NYSIIS [3] codes. For details on the CMDB data fields needed for death clearing the CCR, consult Grabowiecki (1997) and Statistics Canada (1997).

Death clearance can be performed at any time on the CCR. However, the most efficient and effective moment for performing death clearance is just after the completion of the Internal Record Linkage module, that identifies and removes any duplicate patient records on the CCR data base.

The death clearance process has been divided into five steps.

■ **Pre-Processing**

In this phase the input data files for death clearance are verified and prepared for the subsequent processing steps. The specific years of CMDB data available to this death clearance cycle are entered into the system. Based upon these years, the cancer patient population from the CCR, and mortality records from the CMDB are selected.

■ **Direct Match (DM)**

The unique *key* to all the death registrations on the CMDB is a combination of three data fields:

- Year of Death
- Province (/Territory/Country) of Death
- Death Registration Number.

These three fields are also found on the CCR patient record. PTCRs can obtain this information by doing their own death clearance, using local provincial/territorial files of death registrations. Patient records having responses for all three *key* fields first pass through a direct match with the CMDB in an attempt to find mortality records with identical common identifiers. If none is found, they next pass through the probabilistic *record linkage* phase, along with those patient records missing one or more of the *key* match fields. For the records that do match, five data items common to both the patient and CMDB records are compared (Sex, Day of Death, Month of Death, Year of Birth, Month of Birth). On both the CCR patient records and matched CMDB records, the responses must be non-missing and identical. If they are not, both the patient and mortality records are free to participate in the record linkage, where they may link together. Matched pairs that pass the comparison successfully are considered to represent the same person; they then will move on to the *post-processing* phase.

■ **Probabilistic Linkage (PL)**

In order to maximise the possibility of successfully linking to the CMDB file, the file of unmatched CCR patient records is *exploded* by creating, for every person, a separate patient record for each

unique Surname, each part of a hyphenated Surname, and the Birth/Maiden Name -- a process similar to the one used to create the CMDB, described in above. NYSIIS codes are generated for all names.

The two files are then passed through the Generalised Record Linkage System (GRLS), and over 20 important fields are compared using a set of 22 rules. Based on the degree of similarity found in the comparisons, weights are assigned, and the CCR-CMDB record pairs with weights above the pre-established threshold are considered to be linked. When patient records link to more than one mortality record, the pair with the highest weight is taken and the other(s) rejected. Similarly, if two or more patients link to the same CMDB record, the pair with the highest weight is selected.

The threshold weight has been set at such a level that the probability of the linked pairs describing the same person is reasonably high; consequently, manual review is not necessary in the linkage phase. At the same time, the threshold has not been positioned too high, in order to avoid discarding too many valid links, and thus reducing the effectiveness of the record linkage process.

The death information of linked CMDB records is posted onto the CCR patient records, overlaying any previously reported data in these fields. The linked pairs and unlinked CCR patient records join the matched pairs in proceeding to the *post processing* phase of death clearance.

#### ■ **Post-Processing**

Essentially, this phase updates the CCR data base with the results from the *match* and *linkage* phases. Also, the results are communicated to the PTCRs for their review, and for input into their own data bases. Before being updated, copies are made of the patient records from the database. This makes it possible to restore them to their pre-death confirmed state should the matches/linkages be judged to be incorrect later by the PTCRs.

#### ■ **Refusal Processing**

Refusals are PTCR decisions, taken after their review of the feedback reports and files generated in the *post processing* phase, that specific matches and linkages are incorrect -- i.e., that the persons described on the CCR patient records are not the same persons to whose death registrations they matched or linked. In this step, the affected patient records have their confirmation of death reversed, and are restored to their pre-death clearance state.

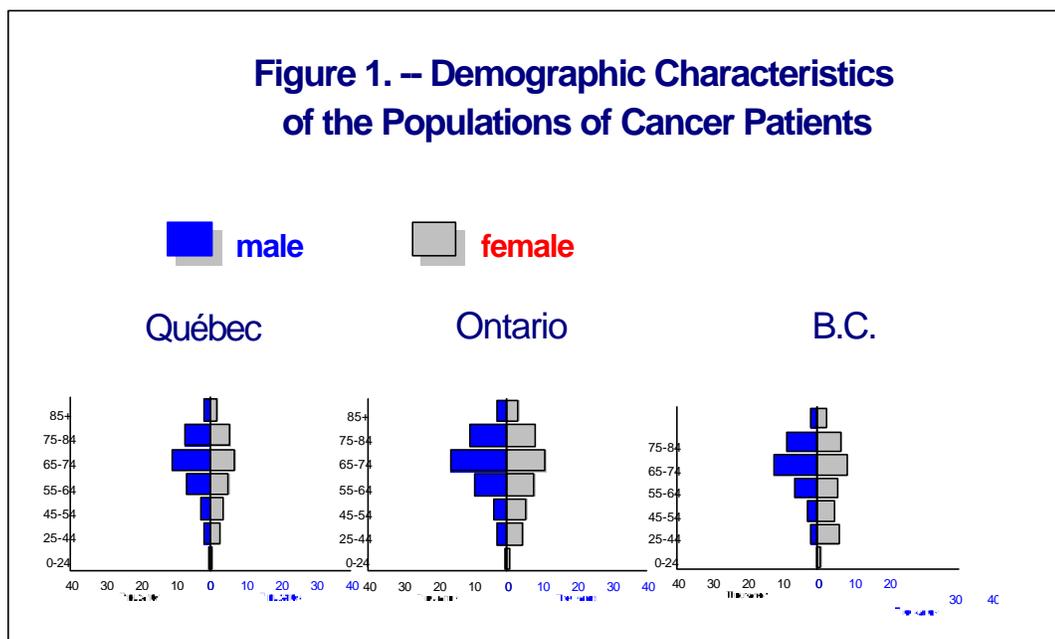
A description of the entire DC Module is available in Grabowiecki (1997) and the detailed specifications of the Direct Match and Probabilistic Linkage can be found in Wysocki and LaBillois (1997).

## Characteristics of the Target Populations for this Study

To perform our comparisons, a subset of the CCR population was selected that could best illustrate the effect of direct match versus probabilistic linkage. Three provinces were chosen: British Columbia, Ontario and Québec. They were picked because they contain, within Canada, the largest populations of cancer patients, and the size of their respective populations is in the same order of magnitude. Québec was specifically taken because its provincial cancer registry does not do death clearance. Consequently, the patient files sent to the CCR by this registry never contain complete death information. Therefore, no cancer patient record from Quebec can obtain a confirmation of death by means of the Direct Match process; all Québec records participate in the Probabilistic Linkage. All other provinces do their own DC, and a significant number of their records on the CCR stand a good chance of being confirmed as dead as a result of the

## Direct Match.

Due to the availability of data from the CCR and the CMDB at this time, we used reference years of diagnosis 1992 and 1993. The distribution by age and sex of the cancer patients in the three provinces is shown in Figure 1, below. It appears that there are only minor differences in the populations of cancer patients between these three provinces. Consequently, such differences are not expected to cause differences in the results of the death clearances.



It is also important to note that the data coming from different provinces are gathered by different PTCRs. Even though there is little difference between them, in terms of coding practices, definitions and timeliness, certain variations still exist. In particular, the data sources used by the PTCRs to build their registries vary considerably among them (Gaudette et al., 1997). These considerations are taken into account in the interpretation of the results.

## Direct Match and Probabilistic Linkage Vs. Only Probabilistic Linkage (Within the Same Province)

### Process

This comparison is done by running the complete DC Module on the CCR data from British Columbia and Ontario. Both the DM and the PL are used to identify pairs for death confirmation. In the second run, any death information contained on the CCR records from these provinces is ignored. The system thus channels all the records directly to the PL. Québec data are not usable for this comparison because of the absence of complete death information on their CCR records. By comparing the two sets of pairs obtained in each approach for death confirmation, it is possible to measure different phenomena:

- overall percentage of accepted pairs (death confirmations) for each approach;
- percentage of pairs that are common to both approaches;

- percentage of pairs that were present in the regular DC process (DM & PL) but not in the PL only;
- percentage of pairs that were not present in the regular DC process (DM & PL) but were found in the PL only; and
- computer time and cost for each approach.

These measures help to evaluate the usefulness of the Direct Match in the DC process and contrarily, the impact of not having the CCR death information previously supplied by PTCRs.

## Results and Observations

The results of this process are summarised in Figure 2, below. When both a DM and PL were performed, the majority of the pairs formed (approximately 95%) came from the DM. This was the case for both of the provinces involved in this part of the study. This result emphasises the importance of high quality death information in effectively matching records on these two files. There can be no direct match unless all of the death fields are identical on the two files, and these account for all but 5% of the total of pairs created in the DM and PL process.

**Figure 2. -- Comparison of Ontario and British Columbia Using Both Methods**

DC Population		DM and PL				PL	
		Matched	Linked	Total	%	Total	%
Ont.	84,926	22,648	1,183	23,831	28.1	23,670	27.9
B.C.	33,103	8,058	360	8,418	25.4	8,367	25.3
Total	118,029	30,706	1,543	32,249	27.3	32,037	27.1

It is evident that in terms of the number of pairs obtained in the end, one can expect little difference between the two methods of death clearance. Additionally, the particular pairs obtained (which specific patients are confirmed) will also be very similar. In this regard, there was less than a 1% difference in the two methods. Those differences that did exist tended to reflect favourably on the DM-PL method. Both methods found the same 32,035 pairs. On a net basis, the DM-PL method found 214 more pairs than did the PL only method. In percentage terms, this represented a negligible amount (again, less than 1%). Of those 214 pairs, roughly 94% were found in the direct match portion of the run; the others were found in the linkage. There were two pairs identified by the linkage-only method and not by its counterpart.

In regard to the actual cost of running the programs under the two different methods, the total for the DM-PL approach was 54% of the total cost incurred in running the PL alone. There is a certain small amount of instability in these numbers since the cost was dependent in part on the level of activity on the mainframe computer at the time that the programs were run. However, the percentage difference in the two costs is substantial even when this is considered. The relatively high cost of the linkage-only approach is due to the fact that the usual preprocessing steps must still be done but, at the same time, the number of records that are compared in the probabilistic linkage is considerably higher than the number used in the DM-PL approach (since many patient records, and their associated death records, will have been accounted for in the DM).

## A Province With Only Probabilistic Linkage Vs. Provinces With Direct Match and Probabilistic Linkage

### Process

For this part, the complete death clearance system is used to process the data of the three selected provinces. It will automatically produce death confirmation pairs by using the Direct Match and the Probabilistic Linkage for British Columbia and Ontario. Simultaneously, it will only apply the Probabilistic Linkage for Québec, because the Québec cancer registry does not report the necessary identifiers for the Direct Match to the CCR. In comparing the death confirmation results obtained for each of the three provinces, it is possible to observe different phenomena. The first is the overall percentage of accepted pairs (death confirmations) for each province, and the possible contrast between Québec and the two others. Another aspect to consider is the comparison of the percentage of death confirmation in Québec versus those obtained with PL only for British Columbia and Ontario in the previous Section. It is also interesting to evaluate the impact of not having the CCR death information previously supplied by PTCRs.

### Results and Observations

The results obtained from the above process are summarised in Figure 3.

**Figure 3. -- Ontario and British Columbia vs. Quebec, Where Only PL Was Possible**

DC Population	DM and PL				PL	
	Matched	Linked	Total	%	Total	%
Qué. 57,252	--	--	--	--	18618	32.5
Ont. 84,926	22,648	1,183	23,831	28.1	--	--
B.C. 33,103	8,058	360	8,418	25.4	--	--

The percentage of pairs found from among the Québec data is rather higher than the corresponding percentages for the other provinces. In addition, all the Québec patient records which contained some death information were successfully linked to a mortality record during probabilistic linkage. This was not the case for all of the Ontario and BC records which contained death information; that is, there were some patients reported as deceased by Ontario and BC which neither matched or linked to a CMDB record. Overall, 32.5% of the Québec records that were in scope were successfully linked to the death file, while 28.1% of the Ontario records and 25.4% of the BC records were matched or linked. As previously noted, the data from Québec does not contain complete death information; it does, however, contain some records where the patient was reported as deceased by this province. It is probable that these were hospital deaths and so it is in turn very unlikely that the corresponding patients are being mistakenly reported as deceased. In essence, these patients can be anticipated to be good candidates to be successfully linked to a death record.

More generally, some cancer patients in Québec receive treatment entirely outside of hospitals and such patients may not then be reported to the CCR. The data from Québec might, therefore, contain a greater proportion of more serious cancers than do the data from the other provinces used in the study. This offers a possible explanation for the higher percentages of cancer patients confirmed in Québec compared to Ontario and B.C.

Finally, we have seen that the differences between the outcomes observed for the Ontario-BC data, using the match and linkage, and the linkage only, in terms of the total number of pairs found, were relatively minor. Again, a greater percentage of pairs were found in Québec than in the other provinces, and possibly because of the reasons outlined above.

## Conclusions

Death Clearance of the CCR using PL only can be conducted with equal effectiveness as the DM-PL approach because of the reporting of high-quality personal and cancer data by the PTCRs. The advantages of the DM-PL method include lower operating costs to perform death clearance (increased efficiency), and greater certainty with the results (minimum manual review of cancer-mortality record pairs by PTCRs).

## Footnotes

- [1] The cancers that are reported to the CCR include all primary, non-benign tumours (with the exception of squamous and basal cell skin cancers, having morphology codes 805 to 808 or 809 to 811, respectively), as well as primary, benign tumours of the brain and central nervous system. In the International Classification of Diseases System – 9<sup>th</sup> Revision (ICD-9), the following codes are included: for benign tumours, 225.0 to 225.9; for *in situ* / intraepithelial / noninfiltrating / noninvasive carcinomas, 230.0 to 234.9; for uncertain and borderline malignancies, 235.0 to 239.9; and finally, for primary site malignancies, 140.0 to 195.8, 199.0, 199.1, and 200.0 to 208.9. Similarly, according to the International Classification of Diseases for Oncology – 2<sup>nd</sup> Edition (ICD-O-2), the target population of cancers includes: all *in situ*, uncertain / borderline, and primary site malignancies (*behaviour codes* 1, 2, or 3), as well as benign tumours (*behaviour code* 0) with topography codes in the range C70.0 to C72.9 (brain and central nervous system).
- [2] Matching entails finding a unique, assigned, identification number on two or more records, thus identifying them as belonging to the same person; whereas linkage concludes that two or more records probably refer to the same person because of the number of similar, personal characteristics found on them.
- [3] NYSIIS (New York State Identification and Intelligence System) assigns the same codes to names that are phonetically similar. It is used to group like-sounding names and thus take into account, during record linkage, variations (and errors) in spelling -- e.g., Burke and Bourque, Jensen and Jonson, Smith and Smythe.

## References

- Gaudette, L.; LaBillois, T.; Gao, R.-N.; and Whittaker, H. (1997). Quality Assurance of the Canadian Cancer Registry, *Symposium 96, Nonsampling Errors*, Proceedings, Ottawa, Statistics Canada.
- Grabowiecki, F. (1997). *Canadian Cancer Registry, Death Clearance Module Overview*, Statistics Canada (internal document).
- Statistics Canada (1994). *Canadian Cancer Registry Data Dictionary*, Health Statistics Division.

Statistics Canada (1997). *Canadian Mortality Data Base Data Dictionary*, Health Statistics Division, (preliminary version).

Wysocki, M. and LaBillois, T. (1997). *Death Clearance Record Linkage Specifications*, Household Survey Methods Division (internal document).

**Note:** For further information, contact: Tony LaBillois, Senior Methodologist, Household Survey Methods Division, Statistics Canada, 16-L, R.H. Coats Building, Ottawa, Ontario K1A 0T6, e-mail: labiton@statcan.ca; Marek Wysocki, Methodologist, Household Survey Methods Division, Statistics Canada, 16-L, R.H. Coats Building, Ottawa, Ontario K1A 0T6, e-mail: wysomar@statcan.ca; Frank Grabowiecki, Project Manager, Health Division, Statistics Canada, 18-H, R.H. Coats Building, Ottawa, Ontario K1A 0T6, e-mail: grabfra@statcan.ca .