Catalogue no. 82-225-XIE — No. 010
ISSN: 1715-2100
ISBN: 978-0-662-46134-0

Canadian Cancer Registry Manuals

Record linkage overview, 2007 edition

by Michel Cormier

Health Statistics Division
Client Custom Services
Main Building, room 2200, 150 Tunney's Pasture Driveway

Telephone: 1-613-951-1746

Statistics Canada Statistique Canada

Canada

Canadian Cancer Registry Manual

Record linkage overview, 2007 edition

By
Michel Cormier

82-225-XIE, no. 010
ISSN: 1715-2100
ISBN: 978-0-662-46134-0

Frequency: occasional

Ottawa

Health Statistics Division

Main Building, room 2200, 150 Tunney's Pasture Driveway
Ottawa, K1A 0T6

How to obtain more information:
Client Custom Services Unit: 1-613-951-1746
E-Mail inquiries: HD-DS@statcan.ca

June 2007

Published by authority of the Minister responsible for Statistics Canada

La version française de cette publication est disponible sur demande (n° 82-225-XIF au catalogue).

Table of contents

1.0 Introduction

The Canadian Cancer Registry (CCR) is comprised of three modules: Core Edit, Internal Record Linkage and Death Clearance. The Core Edit module builds and maintains the CCR. It validates and accepts Provincial/Territorial Cancer Registries (PTCR) data submissions through a comprehensive set of edits, and subsequently posts, updates or deletes information on the CCR database. The Internal Record Linkage module identifies and eliminates duplicate records that may have been loaded onto the database as a result of a name change, a subsequent diagnosis or relocation to another community, province or territory. Finally, the CCR Death Clearance module completes the information on cancer patients by furnishing the Underlying Cause of Death, the Province/Territory/ Country of Death, the Date of Death, and the Death Registration Number. This function is done by direct match and probabilistic linkage of patient records on the CCR to death registrations on the Canadian Mortality Data Base created at Statistics Canada from the annual files of the Canadian Vital Statistics Data Base which is created from provincial and territorial death records.

The CCR Record Linkage module consists of a series of steps applied to the records on the CCR database at a fixed point in time. The goal is to identify groups of patient records, along with their associated tumours, which, although currently registered on the CCR separately, have a high probability of pertaining to the same person or the same tumour. In Record Linkage terminology, this is known as an "internal record linkage" or "unduplication", since the search for duplicate records takes place within a single file.

Obvious duplicates, such as the submission by a PTCR of a P or T record identical to one already posted on the database, will be rejected by the CCR Core Edit module. However, the CCR Core Edit module will not detect more subtle duplicates, such as the same patient submitted at different times under different Patient Identification Numbers (PIN), either by the same PTCR or by different PTCRs. It is precisely these latter cases that the Record Linkage module is designed to identify.

This document provides an overview of the CCR Record Linkage module. For more information, see the CCR Report *User Guide to Record Linkage Feedback Reports C1 and C2*.

2.0 Overview of the CCR record linkage process

The following steps are used in the Record Linkage process:

1) preparations for linkage;
2) pre-processing;
3) record linkage;
4) post-processing;
5) analysis and resolution of groups by PTCRs;
6) resolution entry;
7) resolution processing.

The seven steps used in the Record Linkage process are described in the following sections. The steps are performed sequentially and form a loop or a cycle, which begins and ends with a valid CCR database.

2.1 Preparation for linkage

The CCR database is frozen before beginning the Record Linkage cycle. No new data or updates will be processed through the Core Edit system throughout the Record Linkage process. Decisions taken by PTCRs regarding the resolution of potential duplicates are applied to the CCR database prior to its unfreezing. The database remains frozen for approximately eight weeks.

Although the database is frozen during the Record Linkage cycle, other CCR functions such as the production of tabulation files or the generation of data quality reports can still be performed.

To ensure the security and integrity of the CCR database, a copy is made once the database is frozen. It is this copy that undergoes further processing during the Record Linkage cycle. This step assures the security of the database, should any problem arise during the Record Linkage cycle.

2.2 Pre-processing

The pre-processing step includes the following: "exploding" the P records based on last names (current surname and/or birth surname); joining the P records with the associated T records; defining existing variables as "character" or "numeric" for Record Linkage purposes; creating new variables (including a phonetic version of the last name); treating records with missing values to ensure their proper interpretation during comparisons; and finally, sorting.

The following example illustrates the effect of exploding and joining. For one patient with both a surname and a birth/maiden surname and two tumours on the CCR database, the "exploding" and "joining" would result in four records:

1) P with surname serving as last name, coupled with T1 information;
2) P with surname serving as last name, coupled with T2 information;
3) P with birth/maiden surname serving as last name, coupled with T1 information;
4) P with birth/maiden surname serving as last name, coupled with T2 information.

If patients have multiple names and/or multiple tumours, then the exploding and joining steps increase the size of the file being handled. If the last names are complex (e.g., hyphenated), then even more records will result. During the exploding phase, flags are created to allow the exploded records to be collapsed later and to reflect the content of the original record.

The basic principle in an internal (or one-file) linkage is that each record is compared to every other record and a decision is taken about whether the two records are duplicates. Because the CCR database is relatively large (and made even bigger by exploding and joining), the basic approach would necessitate an enormous number of individual comparisons. In order to make the linkage of large files viable, blocking is usually used. Thus, records are assigned to a block and are subsequently compared only to records in their own block. For the CCR database, the common approach of blocking by a phonetic version of the last name is used. The New York State Identification and Intelligence System (NYSIIS) is used to create a new variable code of the last name. Since the NYSIIS coding results in similar sounding surnames being assigned the same code, patients will only be compared to other patients with similar sounding last names.

As a result of the exploding, an "original" patient record on the CCR database may end up in more than one block if the surname, the birth/maiden surname and/or the two parts of a hyphenated surname are dissimilar enough. This allows these records to be compared with a greater number of records while seeking possible matches. Although blocking precludes many comparisons to increase efficiency, exploding and joining optimises the process and ensures that the best comparisons are done. Thus, a reasonable degree of efficiency is obtained via blocking, without severely limiting the search for matches via exploding.

2.3 Record linkage

Once the CCR database has been pre-processed, the file is then linked. Pairs of records, within each block, are compared according to specified rules based on the contents of selected data fields and weights (numerical scores) are assigned to pairs of records.

There are two methodological techniques used in making the comparisons. The first approach is often referred to as deterministic matching. Deterministic matching produces a match only if all the fields being compared are identical. Although deterministic matching is simple and fast, some valid links will be missed due to minor variations in the fields or missing information.

The second matching technique used on the CCR database for making comparisons is called probabilistic linkage. Probabilistic linkage is a sophisticated method that associates a weight with each field comparison rather than simply allowing two outcomes (identical or not identical). The weight reflects the degree of agreement between the values of the fields on the two records. Outcomes can range from total (exact) agreement to total disagreement, with various levels of partial agreement in between. Close agreement produces a high weight assigned to the outcome. The weights assigned for the field comparisons are summed to obtain a total weight for the pair of records. Usually, more the records in a pair are similar, higher are the total weight for that pair. Record pairs with sufficiently high weights (called linked pairs) are assembled into groups. For example, record A paired with record B may have a high score and record B paired with record C may also have a high score. In this case, these two pairs A, B and B, C would form a group containing A, B and C. Lastly, groups are examined and "unexploded" so that records in the group are restored to the original form they had on the CCR database (in the above example, record C may have in fact been an exploded form of record A).

Thus, the linkage leads to the creation of a set of groups. Since the groups are formed from strongly linked record pairs, the patient records in each group are considered to have a very good chance of representing a single patient. These sets of groups proceed to the next step.

The probabilistic linkage in the CCR Record Linkage module is accomplished using the CANLINK software, developed at Statistics Canada. This package has many built-in features for doing the comparisons and forming the groups. For example, comparison rules can be set up to anticipate certain types of common problems, such as inversion of first and second given names or of months and days in dates. Although the comparisons may be numerous and complex, the CANLINK software at Statistics Canada executes these tasks easily, efficiently and automatically, once the application is set up. CANLINK's functions allows many

simple and complex rules, while more complex, specialized, rules may be added by including user-written PL/1 code, which CANLINK can use during the linkage.

2.4 Post-processing

Once the groups have been identified, a sequence of events aimed at resolving the group created in the Record Linkage phase begins. In essence, the resolution of a group in the post-processing phase consists of one of two choices. If the PTCRs agree that the records in a group do in fact refer to a single patient, then the records will be "merged" and one patient record and some (or all) of the tumour records in the group will be retained on the CCR database, while the others will be deleted. There must agreement on which records to retain and the retained records must obey all relevant CCR edits. If the PTCRs cannot agree that the records in a group refer to a single patient, then the records will remain on the CCR database as originally reported prior to the start of the Record Linkage cycle.

In the post-processing step, the groups are first examined and a tentative merge is identified, consisting of one patient record and some (or all) of the tumour records in the group. This merge is determined according to a pre-defined set of steps designed to yield a logical set of records to be retained. Once constructed, the tentative merge is subjected to the appropriate CCR edits, i.e., the P versus T records, and T versus T records correlation edits. If the tentative merge passes these edits, it becomes the CCR-proposed resolution. If any edit fails, then the merge is not valid and therefore cannot be the resolution of the group. In this case, the CCR-proposed resolution is to leave the records in the group as they were on the database before the Record Linkage cycle began. It is important to note that, at this point, no modification has been brought to the CCR database and a proposal only has been generated.

Another important function of the post-processing step is the generation of a detailed feedback report showing all of the individual records in the group (Record Linkage Feedback Report C1) for each group. The C1 Report indicates the CCR-proposed resolution for each of the groups. The C1 Report is sent to all PTCRs that own a record in the group for review.

2.5 Analysis and resolution of groups by the PTCRs

PTCRs have 3 weeks in which to review the C1 Reports for their groups. All review and any necessary discussion with other PTCRs must take place during this period. In addition, one of the PTCRs that own a P record in the group is designated as the lead registry for reporting purposes. The lead registry must complete the feedback report and return it to the CCR within the review period. The lead registry designation is for administrative purposes and is intended to prevent ambiguity, since only one report per group will be returned to the CCR.

The Record Linkage Feedback Report C1has been designed to serve as a reply form. Once the PTCRs have agreed on a CCR-proposed resolution for the group, the lead registry simply indicates the option that has been selected and sends a copy of the form to the CCR. A space is provided on the C1 Report to indicate an alternate proposal if the PTCRs, involved in the group, have agreed on an alternate way of merging the records. For more detailed information on this step, refer to CCR Report *User Guide to Record Linkage Feedback Reports C1 and C2*.

2.6 Resolution entry

When the completed C1 Reports are returned to the CCR, the decisions taken by the PTCRs are entered into the Record Linkage module via the resolution entry step. The Resolution Entry System (RES) is a specially designed microcomputer software system that allows for key entry of the resolution indicated on the returned C1 Reports. The RES performs some basic validity checks on the inputs (i.e., the resolution decisions sent in by the lead registries), and prepares a file for the next step in the CCR Record Linkage module, for further processing. The RES has built-in defaults for cases not received within the time limit or for

problems encountered during key-enter. When all resolution decisions have been entered, the final step (resolution processing) in the Record Linkage cycle begins.

2.7 Resolution processing

The last step, resolution processing, consists of uniting the file of CCR-proposed resolutions, created in the post-processing step, with the file of resolutions from the RES. The information on the two files is amalgamated and is used to update the CCR database. The CCR-proposed resolution will be implemented, an alternate resolution proposed by the PTCRs will be implemented or the records in the group will be left on the CCR database as they were before linkage.

The execution of alternate resolutions proposed by the PTCRs takes place after editing to ensure that all CCR specifications are met by the proposal. The proposed new combination is edited to ensure that valid relationships are maintained between P versus T records, and T versus T records, similar to merges during post-processing (see section 2.4).

The final activities of resolution processing are: to post the newly resolved records on the database which replace the old, now outdated, records; to unfreeze the CCR database so that regular operations may resume; and, to print confirmation reports (Record Linkage Feedback Report C2) for each group. The C2 Report indicates the action taken during the Record Linkage cycle for each record of each group, including CCRID changes. The C2 Report is sent to the PTCRs that were sent the C1 Report for the particular group.

3.0 Summary

The CCR Record Linkage is initiated independently by the Manager of the CCR through the Record Linkage module, unlike the CCR Core Edit module which is run following the receipt of data submission files (P and T input records) submitted by the PTCRs.

The Record Linkage cycle of the CCR system consists of a series of steps. The initial steps, which are performed automatically, uncover potential duplicate patients, create a proposal for the resolution of duplicate patient records and generate reports. The intermediate steps of the Record Linkage module are not automatic and require an exchange of information between the CCR and the PTCRs. The PTCR input is required to complete the resolution of the suspected duplicate registrations on the CCR database. After the PTCR input is received by the CCR, the resolution decisions are implemented, the database is updated and the reports are generated. Lastly, the database is unfrozen and regular CCR operations resume.

The CCR database is "frozen" during the Record Linkage cycle (approximately eight weeks). The regular updates and CCR Record Linkage are separate activities and are never conducted simultaneously.

The Record Linkage cycle is conducted annually, at a pre-specified time determined by the Canadian Council of Cancer Registries. However, it is possible to operate the Record Linkage cycle more frequently.