Catalogue no. 82-225-XIE — No. 005
ISSN: 1715-2100
ISBN: 0-662-41812-3

Canadian Cancer Registry Manuals

Record Linkage Overview

by Michel Cormier

Health Statistics Division
Client Custom Services
Room 2200, Main Building, Ottawa, K1A 0T6

Telephone: 1 613 951-1746

Statistics Canada Statistique Canada

Canada

Canadian Cancer Registry Manuals

Record Linkage Overview

by

Michel Cormier

82-225-XIE No. 005
ISSN: 1715-2100
ISBN: 0-662-41812-3

Health Statistics Division
Room 2200, Main Building, Ottawa, K1A 0T6

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

October 2005

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

Page

1.0 Introduction 4

2.0 Overview of the CCR record linkage process 4
2.1 Preparations for linkage 5
2.2 Pre-processing 5
2.3 Record linkage 6
2.4 Post-processing 7
2.5 Analysis and resolution of groups by the PTCRs 8
2.6 Resolution entry 8
2.7 Resolution processing 9

3.0 Summary 9

1.0 Introduction

The primary function of the Core Edit system of the Canadian Cancer Registry (CCR) is to accept patient (P) and tumour (T) records from the Provincial/Territorial Cancer Registries (PTCRs), run them through a comprehensive set of edits and post the valid data to the CCR database. The Core Edit system also produces reports that summarize the results of the process (including any new CCRID numbers issued) and provides detailed feedback on the records that were not posted on the database. The Record Linkage module of the CCR is another important function, which acts upon the database to identify possible duplicate records. While the Record Linkage function is an integral part of the CCR system, it is a distinct processing function.

The CCR Record Linkage module consists of a series of steps applied to the records on the CCR database at a given point in time. The goal is to find 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. 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 due to the edits in 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 (PINs), 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 process used in the Record Linkage module consists of the following steps:

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.

These steps, described in the following sections, are performed sequentially and form a loop or a cycle, which begins and ends with a valid CCR database.

2.1 Preparations for linkage

To begin the Record Linkage cycle the CCR database is frozen. During this time, no data will be processed through the Core Edit system. The CCR database must be stable throughout the Record Linkage cycle, since PTCR decisions relating to the resolution of potential duplicates will be based on the information it contained at the beginning of the cycle. The period during which the database is frozen lasts approximately six weeks.

Although the database is frozen in terms of processes which would alter it, other CCR functions such as production of tabulation files or generation of data quality reports can be performed.

Once the CCR database is frozen, an exact copy is made. It is this copy which 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 involves "exploding" the P records based on last names (current surname and/or birth surname), joining P records with the associated T records, defining existing variables as "character" or "numeric" for Record Linkage purposes, creating some new variables (including a phonetic version of the last name), treating records with missing values, to ensure their proper interpretation during comparisons, and finally, doing some sorting.

The following example illustrates the effect of exploding and joining. If there existed one patient with both a surname and a birth/maiden surname and two tumours on the CCR database, then 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.

Clearly, if any of the patients have multiple names and/or multiple tumours, then the exploding and joining step increases the size of the file being handled. If the last names are more complex (e.g., hyphenated), then even more records will result. During the exploding, flags are created to allow exploded records to be collapsed later, to properly 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. Since the CCR database is large to begin with (and made even bigger by the exploding and joining), using this basic approach would involve a tremendous number of individual comparisons. In order to make the linkage of large files viable, blocking is usually used. In other words, records are assigned to a block, and are subsequently compared only to records in their own block. Clearly, the information used to assign the record to a block should be chosen in order that, for a given record, the chance of finding a match within the block is definitely higher than the chance of finding a match with a record outside the block. For the CCR, the common approach of blocking by a phonetic version of the last name is used. This is accomplished by creating a new variable, which is the NYSIIS (New York State Identification and Intelligence System) code of the last name. The NYSIIS coding results in similar sounding surnames being assigned the same code. Therefore, patients will only be compared to other patients with similar sounding last names (i.e., only to others patients within their block).

Blocking will admittedly preclude many comparisons in the name of efficiency, but the previous step of exploding and joining ensures that the important comparisons are done. As a result of the exploding, a given "original" patient record on the CCR 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. Thus, combining the blocking with the exploding and joining optimises the process in the sense that a reasonable degree of efficiency is obtained (via blocking), without limiting too severely the search for matches (via exploding).

2.3 Record linkage

Once pre-processed, the file is then linked. In other words, within each block, pairs of records 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 basically two approaches to making the comparisons. One approach is often called deterministic matching. It consists of achieving a match only if the fields being compared are identical. This is a simple and fast approach, but clearly some valid links will be missed if there are even minor variations in the fields.

The second approach, used in the CCR, is called probabilistic linking. This is a sophisticated method, since rather than simply allowing two outcomes (identical or not), a weight is associated with each field comparison. The weight depends on 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. For a given comparison the closer the agreement, the higher the weight assigned to the outcome. The weights assigned for the field comparisons are then summed to obtain a total weight for the pair of records. Generally, the more the records in a pair are similar, the higher the total weight for that pair. Record pairs with sufficiently high weights (called linked pairs) continue in the process. The linked pairs are next assembled into groups; for example, record

A paired with B may have a high score and B paired with 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. Finally, 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, C may have in fact been an exploded form of A).

Thus, the linkage leads to the creation of a set of groups. Since they 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. The set of groups passes on 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 allow to establish 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, the post-processing phase begins a sequence of steps aimed at resolving the group. Essentially, the resolution of a group will take one of the two following forms. If it is agreed that the records in a group do in fact refer to a single patient, they will be "merged". This means that one patient record and some (or all) of the tumour records in the group are retained, while the others are deleted. In order for the resolution of a group to be a merge, there must be agreement on which records to retain, and the retained records must obey all relevant CCR edits. Otherwise, the resolution will be to keep all original records in the group on the CCR database, as they were 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 would be 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, a proposal has only been generated.

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

2.5 Analysis and resolution of groups by the PTCRs

PTCRs usually have 30 days in which to review the C1 Reports for their groups. All review and any necessary discussions with other PTCRs must take place during this period. In addition, one of the PTCRs which owns 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 an administrative purpose and aims to prevent ambiguity, since only one report per group will be returned to the CCR.

The feedback report has been designed to serve as a reply form. Once the PTCRs have agreed on a resolution for the group, the lead registry simply indicates the number of the option that has been selected and sends a copy of the form to the CCR. In the case where the CCR-proposed resolution is rejected by the PTCRs involved in the group, and where they have agreed on an alternate way of merging the records, a space is provided on the C1 Report to write the alternate proposal. In formulating the reply, lead registries consult with the other PTCRs owning records in the group and are therefore in a position to reply on behalf of the other PTCRs. 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 by the PTCRs, the decisions they contain are entered into the Record Linkage module via the resolution entry step. The Resolution Entry System (RES) is a specially designed microcomputer software system to allow for key entry of the resolution decision indicated on the returned C1 Reports. The RES begins with a file of information similar to the one shown on the feedback reports sent to registries. It first performs some basic validity checks on the inputs (i.e., the resolution decisions sent in by the lead registries), and then prepares a file for transfer to the next step in the CCR Record Linkage module, for further processing. The RES has built-in defaults to handle cases for which the feedback report for a group was not received within the time limit, from the lead registry, or those where problems were encountered while attempting to key-enter the resolution response for a group. When all resolution decisions have been entered, the final step 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 these two files is amalgamated and is used to update the CCR database. Depending on the resolution option selected, 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.

In the case of alternate resolutions proposed by the PTCRs, execution is not automatic, but takes place only after editing, to ensure that all CCR specifications are met by the proposal. For example, an alternate merge may have been specified, that would unite a patient with a tumour that was not, previously, associated with that patient. The new combination will be edited to ensure that valid relationships are maintained between P versus T records, and T versus T records, just as was done for each tentative merge constructed by the CCR during post-processing (see section 2.4).

The final activities of this step are to post the newly resolved records on the database in order to replace their old, now outdated, records, to unfreeze the CCR database so that regular operations may resume, and to print a confirmation report (Record Linkage Feedback Report C2) for each group. This confirmation report indicates the action taken as a result of the Record Linkage cycle, on each record, of each group, including any CCRID changes. A copy of the C2 Report is sent to each PTCR that received a C1 Report for the particular group.

3.0 Summary

The CCR Record Linkage cycle is initiated by the CCR, as opposed to the CCR Core Edit module which is triggered by the receipt of files of P and T records submitted by the PTCRs.

The Record Linkage module of the CCR system consists of a series of steps accomplished during the Record Linkage cycle. The initial steps uncover potential duplicate patients, create a proposal for the resolution of these duplicate patient records and generate reports. These steps are automatically performed at the CCR. However, the PTCR input is required to complete the resolution of the suspected duplicate registrations on the CCR database. Thus, the intermediate steps of the Record Linkage module are not automatic. It depends on an exchange of information between the CCR and the PTCRs. After the PTCR input is received by the CCR, the automatic process resumes and the resolution decisions are implemented, the database is updated and further reports are generated. Finally, the database is unfrozen and normal CCR operations resume.

The CCR database is "frozen" during the Record Linkage cycle (approximately six weeks). The regular/correction and Record Linkage cycles are therefore separate activities and are never operated simultaneously.

The Record Linkage cycle is conducted annually, at a pre-specified time. It is possible to operate the Record Linkage cycle more frequently.

During the Record Linkage cycle, records are matched using the information they contain at the time the CCR database is frozen. Information provided by PTCRs in subsequent regular or correction submissions will be taken into account during the next Record Linkage cycle.