Nº 82-225-XIF au catalogue — Nº 005
ISSN: 1715-2119
ISBN: 0-662-70368-5

Manuels de procédures du registre canadien du cancer

Aperçu du couplage des enregistrements

par Michel Cormier

Division de la statistique de la santé
Services personnalisés à la clientèle
Immeuble principal, pièce 2200, Ottawa, K1A 0T6

Telephone: 1 613 951-1746

Statistique Canada Statistics Canada

Canada

Manuels de procédures du registre canadien du cancer

Aperçu du couplage des enregistrements

par

Michel Cormier

82-225-XIF No. 005
ISSN: 1715-2119
ISBN: 0-662-70368-5

Division de la statistique de la santé
Pièce 2200, Immeuble principal, Ottawa, K1A 0T6

Pour obtenir plus d'information :
Services personnalisés à la clientele : 1 613 951-1746
Courriel: HD-DS@statcan.ca

Octobre 2005

Publication autorisée par le ministre responsable de Statistique Canada

This publication is available in English upon request (catalogue no. 82-225-XIE*).*

Table des matières

Page

1.0 Introduction

La fonction première du module principal de contrôle du Registre canadien du cancer (RCC) est d'accepter les enregistrements patient (P) et tumeur (T) présentés par les registres provinciaux ou territoriaux du cancer (RPTC), de les soumettre à un ensemble exhaustif de contrôles et d'inscrire les données valides dans la base de données du RCC. Le module principal de contrôle produit aussi des rapports qui résument les résultats du processus (y compris tous les nouveaux numéros d'identification du RCC attribués) et fournit une rétroaction détaillée sur les enregistrements qui n'ont pas été inscrits dans la base de données. Le module de couplage des enregistrements du RCC est une autre fonction importante, qui permet d'examiner la base de données pour repérer les enregistrements peut-être en double. Bien que la fonction de couplage des enregistrements fasse partie intégrante du système du RCC, elle est une fonction de traitement distincte.

Le module de couplage des enregistrements du RCC consiste en une série d'étapes par lesquelles passent les enregistrements de la base de données à un moment précis. L'objectif consiste à repérer des groupes d'enregistrements patient, avec les enregistrements tumeur qui leurs sont associés, qui, bien qu'inscrits séparément dans le RCC, se rapportent fort probablement à la même personne. Dans le jargon du couplage des enregistrements, ceci est désigné comme étant le « couplage interne des enregistrements » ou l'« élimination des doubles », puisque la recherche des enregistrements en double a lieu au sein d'un seul fichier.

Les enregistrements manifestement en double, tels que la soumission par un RPTC d'un enregistrement P ou T identique à un enregistrement figurant déjà dans la base de données, seront rejetés grâce aux contrôles effectués par le module principal de contrôle du RCC. Toutefois, le module principal de contrôle du RCC ne décèlera pas les enregistrements en double moins évidents, tels qu'un même patient soumis à des moments différents, sous des numéros d'identification du patient (NIP) distincts, soit par le même RPTC ou par des RPTC différents. Ce sont précisément les cas de ce genre que le module de couplage des enregistrements a été créé pour repérer.

Ce document donne un aperçu du module de couplage des enregistrements du RCC. Bien qu'une connaissance approfondie du système du RCC ne soit pas nécessaire pour la compréhension de ce rapport, quelques notions de base pourraient être utiles lors de la lecture de cet aperçu. Pour de plus amples informations, consulter le rapport du RCC *Guide d'utilisation des rapports de rétroaction C1 et C2 du couplage des enregistrements*.

2.0 Aperçu du processus de couplage des enregistrements du RCC

Le processus utilisé dans le module de couplage des enregistrements est constitué des étapes suivantes :

1) préparatifs avant le couplage;
2) pré-traitement;
3) couplage des enregistrements;
4) post-traitement;
5) analyse des groupes et choix d'une solution par les RPTC;
6) entrée des solutions;
7) traitement des solutions.

Ces étapes, décrites dans les sections suivantes, sont exécutées séquentiellement et forment une boucle ou un cycle, qui a pour origine et pour point final une base de données du RCC valide.

2.1 Préparatifs avant le couplage

Pour entamer le cycle de couplage des enregistrements, la base de données du RCC est figée. Durant cette période, aucune donnée ne sera traitée par le module principal de contrôle. La base de données du RCC doit demeurer stable durant le cycle de couplage des enregistrements, puisque les décisions adoptées par les RPTC concernant la résolution des doubles potentiels seront basées sur l'information qu'elle contenait au début du cycle. La période pendant laquelle la base de données est figée dure environ six semaines.

Bien que la base de données du RCC est figée en termes de processus susceptibles de la modifier, d'autres fonctions du RCC, telles que la production de fichiers de classement ou de rapports sur la qualité des données peuvent être exécutées.

Lorsque la base de données du RCC est figée, une copie exacte en est produite. C'est sur cette copie que seront exécutés les traitements durant le cycle de couplage des enregistrements. Cette étape vise à protéger la base de données, au cas où un problème surviendrait durant le cycle de couplage des enregistrements.

2.2 Pré-traitement

Le pré-traitement consiste à « ventiler » les enregistrements P en se basant sur le nom de famille (nom de famille actuel et/ou nom de famille à la naissance), à unir les enregistrements P avec les enregistrements T associés, à définir les variables existantes comme étant de type « caractère » ou « numérique » aux fins du couplage des enregistrements, à créer quelques nouvelles variables (incluant une version phonétique du nom de famille), à traiter les enregistrements dont certaines valeurs manquent, pour

s'assurer qu'ils soient interprétés correctement durant les comparaisons et enfin, à effectuer certains tris.

L'exemple suivant illustre l'effet de la ventilation et de l'union. S'il existait un patient ayant un nom de famille ainsi qu'un nom de famille à la naissance / nom de jeune fille et deux tumeurs dans la base de données du RCC, alors la « ventilation » et l'« union » résulteraient en quatre enregistrements :

1) P avec le nom de famille servant de nom de famille, uni à l'information T1;
2) P avec le nom de famille servant de nom de famille, uni à l'information T2;
3) P avec le nom de famille à la naissance / nom de jeune fille servant de nom de famille, uni à l'information T1;
4) P avec le nom de famille à la naissance / nom de jeune fille servant de nom de famille, uni à l'information T2.

Manifestement, si un des patients possède plusieurs noms et/ou est atteint de plusieurs tumeurs, alors l'étape de la ventilation et de l'union augmente la taille du fichier en cours de traitement. Si les noms de famille sont plus complexes (p.ex., noms composés), alors un nombre encore plus grand d'enregistrements en résultera. Durant la ventilation, des indicateurs sont créés en vue de permettre le regroupement futur des enregistrements ventilés, afin de refléter correctement le contenu de l'enregistrement original.

Le principe fondamental d'un couplage interne (ou au sein d'un fichier unique) est que chaque enregistrement est comparé à tous les autres et une décision est prise à savoir si les deux enregistrements sont en double. Étant donné que la base de données du RCC est vaste au départ (et est encore agrandie par la ventilation et l'union), l'application de cette méthode de base entraînerait une énorme quantité de comparaisons individuelles. Afin de rendre le couplage de grands fichiers viable, un groupage est généralement utilisé. En d'autres mots, les enregistrements sont assignés à un bloc, puis sont comparés uniquement aux enregistrements qui forment leur bloc. De toute évidence, l'information utilisée pour assigner un enregistrement à un bloc devrait être choisie de telle sorte que, pour un enregistrement donné, la probabilité de découvrir un appariement au sein du bloc soit nettement plus élevée que celle de repérer un appariement avec un enregistrement non compris dans le bloc. Dans le cas du RCC, la méthode courante de groupage d'après une version phonétique du nom de famille est utilisée. Ceci est accompli grâce à la création d'une nouvelle variable, à savoir le code NYSIIS (New York State Identification and Intelligence System) du nom de famille. Le codage du NYSIIS a pour résultat que les noms de famille ayant la même prononciation reçoivent le même code. Donc, les patients ne seront comparés qu'aux autres patients ayant un nom de famille qui se prononce d'une manière semblable (c.-à-d., uniquement aux autres patients appartenant au même bloc).

Le groupage exclut certes de nombreuses comparaisons au nom de l'efficacité, mais l'étape précédente de ventilation et d'union assure que les comparaisons importantes soient effectuées. En effet, à la suite de la ventilation, un enregistrement patient « original » donné dans le RCC peut se

retrouver dans plus d'un bloc (si le nom de famille actuel, le nom de famille à la naissance / nom de jeune fille et/ou les deux éléments d'un nom de famille composé sont suffisamment différents). Ceci permet à ces enregistrements d'être comparés à un plus grand nombre d'enregistrements pendant la recherche d'appariements éventuels. Donc, combiner le groupage à la ventilation et l'union optimise le processus, en ce sens qu'un degré raisonnable d'efficacité est atteint (grâce au groupage), sans toutefois restreindre trop sévèrement la recherche d'appariements (grâce à la ventilation).

2.3 Couplage des enregistrements

Une fois pré-traité, le fichier est alors soumis au couplage. Autrement dit, au sein de chaque bloc, les paires d'enregistrements sont comparées selon des règles précises, en se basant sur le contenu de zones de données sélectionnées, puis des facteurs de pondération (cotes numériques) sont attribués aux paires d'enregistrements.

Il y a fondamentalement deux méthodes pour effectuer les comparaisons. L'une des méthodes est souvent appelée appariement déterministe. Elle consiste à ne faire un appariement que si les zones comparées sont identiques. Cette méthode est simple et rapide, mais manifestement certains couplages valides seront omis en cas de discordances, même mineures, entre les zones.

La deuxième méthode, utilisée par le RCC, porte le nom de couplage probabiliste. Il s'agit d'une méthode sophistiquée, puisqu'au lieu d'accepter simplement deux résultats (identiques ou non identiques), un facteur de pondération est attribué à chaque comparaison de zones. Le facteur de pondération dépend du degré de concordance entre les valeurs qui apparaissent dans les zones des deux enregistrements. Les résultats peuvent varier de la concordance totale (exacte) à la discordance totale, en passant par divers degrés de concordance partielle. Pour une comparaison donnée, le facteur de pondération attribué au résultat est d'autant plus élevé que la concordance est forte. Les facteurs de pondération attribués aux comparaisons de zones sont ensuite additionnés pour obtenir un facteur de pondération total pour la paire d'enregistrements. En général, plus les enregistrements d'une paire sont semblables, plus le facteur de pondération total pour cette paire sera élevé. Les paires d'enregistrements dont les facteurs de pondération sont suffisamment élevés (appelées paires couplées) sont soumises aux autres étapes du processus. Les paires couplées sont ensuite assemblées en groupes; par exemple, l'enregistrement A apparié avec le B peuvent avoir une cote élevée, et B apparié à C peuvent aussi avoir une cote élevée. Le cas échéant, ces deux paires A, B et B, C formeraient un groupe contenant A, B et C. Enfin, les groupes sont examinés, puis « dé-ventilés » de façon à ce que les enregistrements dans le groupe retrouvent la forme originale qu'ils avaient dans la base de données du RCC (dans l'exemple susmentionné, C pourrait en fait être une forme ventilée de A).

Donc, le couplage conduit à la formation d'un ensemble de groupes. Puisque ces derniers sont constitués de paires d'enregistrements fortement liés, les enregistrements patient de chaque groupe ont de fortes chances de représenter un seul patient. L'ensemble des groupes passe alors à l'étape suivante.

Le couplage probabiliste du module de couplage des enregistrements du RCC est exécuté au moyen du logiciel CANLINK, mis au point par Statistique Canada. Ce logiciel comprend de nombreuses fonctions intégrées qui permettent d'effectuer des comparaisons et de former des groupes. Par exemple, les règles de comparaison peuvent être établies avec l'objectif de prévenir certains types de problèmes courants, tels que l'inversion du premier et du deuxième prénoms, ou celle des mois et des jours dans les dates. Bien que les comparaisons peuvent être nombreuses et complexes, le logiciel CANLINK de Statistique Canada exécute ces tâches facilement, efficacement et automatiquement, une fois l'application installée. Les fonctions de CANLINK permettent d'établir nombre de règles simples ou complexes, tandis que des règles spécialisées, plus complexes, peuvent être ajoutées grâce à l'inclusion du code PL/1 inscrit par l'utilisateur, que CANLINK pourra utiliser durant le couplage.

2.4 Post-traitement

Une fois les groupes repérés, la phase du post-traitement entame une série d'étapes ayant pour but de résoudre chaque groupe. Essentiellement, la solution adoptée pour un groupe prendra l'une des deux formes suivantes. S'il est convenu que les enregistrements d'un groupe se rapportent effectivement à un seul patient, ils seront « fusionnés ». Ceci signifie qu'un des enregistrements patient et certains enregistrements tumeur (voire tous) du groupe sont retenus, tandis que les autres sont supprimés. Afin que la solution adoptée pour un groupe soit une fusion, il doit y avoir un consensus sur les enregistrements à retenir, et les enregistrements retenus doivent réussir tous les contrôles pertinents du RCC. Autrement, la solution adoptée consiste à garder tous les enregistrements originaux du groupe dans la base de données du RCC, tels qu'ils y figuraient avant le début du cycle de couplage des enregistrements.

À l'étape du post-traitement, les groupes sont d'abord examinés et une proposition de fusion est identifiée, englobant un des enregistrements patient et certains enregistrements tumeur (voire tous) du groupe. Cette fusion est déterminée en suivant une série prédéfinie d'étapes conçues de façon à produire un ensemble logique d'enregistrements à retenir. Une fois élaborée, la proposition de fusion est soumise aux contrôles pertinents du RCC, c.-à-d., les contrôles de cohérence entre enregistrements P et T, et entre enregistrements T et T. Si la proposition de fusion n'est rejetée lors d'aucun de ces contrôles, elle devient la solution proposée par le RCC. En revanche, si elle est rejetée lors d'un des contrôles, la fusion est invalide et ne peut donc pas être la solution choisie pour le groupe. Dans ce cas, la solution proposée par le RCC consiste à garder les enregistrements du groupe tels qu'ils figuraient dans la base de données avant le début du cycle de couplage des

enregistrements. Il est important de noter qu'à ce stade, aucune modification n'a été apportée à la base de données du RCC, une proposition a uniquement été élaborée.

Une autre fonction importante de l'étape du post-traitement est la production d'un rapport de rétroaction détaillé (Rapport de rétroaction C1 du couplage des enregistrements) pour chaque groupe, précisant tous les enregistrements individuels du groupe. Le rapport indique la solution proposée par le RCC pour chacun des groupes. Le Rapport C1 est envoyé pour révision à tous les RPTC auxquels appartiennent un des enregistrements du groupe.

2.5 Analyse des groupes et choix d'une solution par les RPTC

Les RPTC disposent habituellement de 30 jours pour examiner les rapports C1 pour leurs groupes. Toutes les révisions et la consultation éventuelle d'autres RPTC doivent avoir lieu durant cette période. De plus, un des RPTC propriétaire d'un enregistrement P du groupe est nommé registre principal aux fins de déclaration. Le registre principal doit compléter le rapport de rétroaction et le renvoyer au RCC à l'intérieur de la période accordée pour la révision. La désignation du registre principal est d'ordre administratif et vise à éviter les ambiguïtés, puisqu'un seul rapport par groupe sera renvoyé au RCC.

Le rapport de rétroaction a été conçu pour servir de formulaire-réponse. Une fois que les RPTC ont convenu de la solution à adopter pour le groupe, le registre principal indique simplement le numéro de l'option choisie et envoie un exemplaire du formulaire au RCC. Dans le cas où la solution proposée par le RCC est rejetée par les RPTC concernés par le groupe, et où ces derniers ont convenu d'une autre façon de fusionner les enregistrements, un espace est réservé sur le rapport C1 pour l'inscription de la solution de rechange. Lors de l'élaboration de la réponse, les registres principaux doivent consulter les autres RPTC propriétaires d'enregistrements qui figurent dans le groupe, et sont ainsi en mesure de répondre au nom des autres RPTC. Pour plus de renseignements sur cette étape, consulter le rapport du RCC *Guide d'utilisation des rapports de rétroaction C1 et C2 du couplage des enregistrements*.

2.6 Entrée des solutions

Lorsque les rapports C1 dûment remplis sont renvoyés au RCC par les RPTC, les décisions qui y sont indiquées sont entrées dans le module de couplage des enregistrements durant l'étape d'entrée des solutions. Le Système d'entrée des solutions (SES) est un système micro logiciel spécialement conçu pour permettre l'entrée manuelle des solutions indiquées sur les rapports C1 renvoyés. Le SES débute avec un fichier d'information similaire à celui qui figure sur les rapports de rétroaction envoyés aux registres. Il effectue d'abord certains contrôles de validité de base sur les entrées (c.-à-d., les solutions communiquées par les registres principaux), puis prépare un fichier pour le transfert à l'étape suivante du module de couplage des enregistrements du RCC, afin de continuer le traitement. Des

options implicites sont intégrées au SES pour traiter les cas où le rapport de rétroaction C1 pour un groupe n'a pas été reçu dans le délai prévu du registre principal, ou ceux où des problèmes surviennent pendant la saisie manuelle de la solution recommandée pour un groupe. Quand toutes les décisions de solutions ont été entrées, l'étape finale du cycle de couplage des enregistrements commence.

2.7 Traitement des solutions

La dernière étape, le traitement des solutions, consiste à unir le fichier des solutions proposées par le RCC, produit à l'étape du post-traitement, au fichier des solutions du SES. L'information de ces deux fichiers est amalgamée et est utilisée pour mettre à jour la base données du RCC. Selon l'option de solution sélectionnée, la solution proposée par le RCC sera mise en oeuvre, la solution de rechange proposée par les RPTC sera mise en oeuvre, ou les enregistrements du groupe demeureront dans la base de données du RCC tels qu'ils y figuraient avant le couplage.

Dans le cas des solutions de rechange proposées par les RPTC, l'exécution n'est pas automatique, mais n'a lieu qu'après le contrôle, pour s'assurer que la proposition est conforme à toutes les spécifications du RCC. Par exemple, une solution de rechange pourrait avoir été spécifiée, qui unirait un patient à une tumeur qui n'était pas, antérieurement, associée à ce patient. La nouvelle combinaison sera contrôlée pour s'assurer que des relations valides persistent entre enregistrements P et T, et entre enregistrements T et T, comme cela a été fait pour chaque proposition de fusion élaborée par le RCC durant le post-traitement (voir la section 2.4).

Les dernières activités de cette étape consistent à inscrire les enregistrements nouvellement résolus dans la base de données afin de remplacer les anciens enregistrements, désormais périmés, à défiger la base de données du RCC afin que les opérations régulières puissent reprendre, et à imprimer un rapport de confirmation (rapport de rétroaction C2 du couplage des enregistrements) pour chaque groupe. Ce rapport de confirmation indique les mesures prises à la suite du cycle de couplage des enregistrements pour chaque enregistrement, de chaque groupe, incluant toute modification à l'ID RCC. Un exemplaire du rapport C2 est envoyé à chaque RPTC qui a reçu un rapport C1 pour le groupe en question.

3.0 Résumé

Le cycle de couplage des enregistrements du RCC est amorcé par le RCC, contrairement au module principal de contrôle du RCC, qui est déclenché par la réception des fichiers d'enregistrements P et T transmis par les RPTC.

Le module de couplage des enregistrements du système du RCC comprend une série d'étapes accomplies durant le cycle de couplage des enregistrements. Les premières étapes consistent à rechercher les enregistrements patient potentiellement en double, à créer une proposition de solution pour ces enregistrements patient en double et à produire des rapports. Ces étapes sont effectuées automatiquement au RCC. Cependant, l'intervention des RPTC est nécessaire pour adopter la solution définitive concernant les enregistrements soupçonnés d'être en double dans la base de données du RCC. Par conséquent, l'exécution des étapes intermédiaires du module de couplage des enregistrements n'est pas automatique. Elle dépend d'un échange d'information entre le RCC et les RPTC. Après la réception par le RCC des commentaires des RPTC, le processus automatique reprend et les solutions adoptées sont mises en oeuvre, la base de données est mise à jour et de nouveaux rapports sont produits. Enfin, la base de données est défigée et les opérations courantes du RCC reprennent.

La base de données du RCC est « figée » durant le cycle de couplage des enregistrements (environ six semaines). Les cycles ordinaire/de correction et du couplage des enregistrements représentent donc des activités distinctes et ne sont jamais exécutées simultanément.

Le cycle de couplage des enregistrements est activé annuellement, à un moment prédéterminé. Il est possible d'exécuter le cycle de couplage des enregistrements plus fréquemment.

Durant le cycle de couplage des enregistrements, les enregistrements sont appariés en utilisant l'information qu'ils contiennent au moment où la base de données du RCC est figée. L'information communiquée par les RPTC durant les soumissions subséquentes, régulière ou de correction, sera prise en considération au cours du cycle suivant de couplage des enregistrements.