Recueil du Symposium 2016 de Statistique Canada
Croissance de l'information statistique : défis et bénéfices

Les défis du jumelage et de l'utilisation de données administratives provenant de sources différentes

Philippe Gamache[1]

Résumé

À l'Institut national de santé publique du Québec, le Système intégré de surveillance des maladies chroniques du Québec (SISMACQ) est utilisé quotidiennement depuis environ quatre ans. Les bénéfices de ce système sont nombreux pour mesurer plus précisément l'ampleur des maladies, pour évaluer adéquatement l'utilisation des services de santé et pour identifier certains groupes à risque. Or, au fil des mois, divers problèmes sont apparus et ont nécessité une réflexion plus ou moins longue et élaborée. Des problèmes touchant différentes sphères d'activité, du jumelage à la qualité des données en passant par la coordination entre de multiples utilisateurs et le respect des obligations légales. Cette présentation a pour objet de décrire les principaux défis associés à l'utilisation des données du SISMACQ, ainsi que de présenter quelques pistes de solution. Sera notamment abordé le traitement de cinq sources de données qui proviennent non seulement de cinq sources différentes, mais dont l'utilité première n'est pas la surveillance des maladies chroniques. La qualité variable des données, autant entre fichiers qu'à l'intérieur d'un même fichier, sera aussi discutée. Certaines situations reliées à l'utilisation simultanée du système par plusieurs utilisateurs seront aussi exposées. Des exemples d'analyses sur de grands ensembles de données ayant causé des tracas seront donnés. Également, quelques défis entourant la divulgation et le respect des ententes légales seront brièvement évoqués.

Mots Clés : données administratives; jumelage; qualité; divulgation; surveillance.

1. Le Système intégré de surveillance des maladies chroniques (SISMACQ)

1.1 Introduction

Depuis plus de 15 ans, l'Institut national de santé publique du Québec (INSPQ) effectue la surveillance des maladies chroniques en vertu d'un mandat confié par le ministère de la Santé et des Services sociaux du Québec (MSSS). Initialement, cette surveillance populationnelle s'effectuait à l'aide de fichiers administratifs (décès, hospitalisations) et des enquêtes, mais ces sources de données étaient traitées de manière indépendante. Depuis quelques années, le jumelage de plusieurs fichiers administratifs a grandement amélioré la fiabilité des estimations des nombreux indicateurs produits à l'INSPQ, en particulier à ce qui a trait à la surveillance des maladies chroniques.

Le système basé sur ce jumelage récent de données issues de fichiers administratifs se nomme le système intégré de surveillance des maladies chroniques du Québec (SISMACQ). À ce jour, le SISMACQ couvre la période d'avril 1996 à mars 2014. Bien que les avantages de ce système soient nombreux, son exploitation est une source de nombreux défis méthodologiques, technologiques et techniques. Comment combiner adéquatement cinq bases de données qui sont non seulement administrées par des entités différentes, mais surtout dont l'utilité première n'est pas la surveillance des maladies chroniques ? Est-il possible de réduire les impacts d'une qualité variable des données ? De multiples statisticiens, analystes, étudiants et stagiaires peuvent-ils exploiter le SISMACQ simultanément sans créer de problèmes ? Quelles stratégies doivent être employées pour faciliter l'exploitation tout en respectant plusieurs contraintes légales ?

1.2 Les différences entre les cinq bases de données du SISMACQ

Cinq principaux fichiers administratifs jumelés forment le SISMACQ. Il s'agit du fichier des décès, du fichier des hospitalisations (Maintenance et exploitation des données pour l'étude de la clientèle hospitalière), du registre des personnes assurées à l'assurance maladie du Québec (FIPA), du fichier des services médicaux rémunérés à l'acte, et enfin du fichier des services pharmaceutiques. Ces cinq bases de données sont jumelées à la source via le numéro

[1]Philippe Gamache, Institut national de santé publique du Québec, 945 avenue Wolfe, Québec, G1V 5B3 philippe.gamache@inspq.qc.ca

d'assurance maladie (NAM), qui est ensuite encrypté de sorte que l'INSPQ reçoit un numéro d'individu unique différent du NAM.

Ces cinq fichiers administratifs n'ont évidemment pas été sélectionnés au hasard, et les intervenants impliqués avaient reconnus leur potentiel à améliorer la surveillance des maladies chroniques. Ainsi, à lui seul, le fichier des hospitalisations permet seulement d'identifier les malades les plus graves, c'est-à-dire ceux dont la gravité de la maladie a mené à au moins un séjour hospitalier. Le combiner aux fichiers des services médicaux et des services pharmaceutiques allait permettre de détecter aussi les malades chroniques utilisant des services de santé autre qu'hospitaliers.

Il n'en demeure pas moins que l'utilité première des cinq bases de données administratives n'était pas du tout la surveillance des maladies chroniques. Qui plus est, l'administrateur de chaque base n'est pas le même, de sorte que la gestion et la validation des données est loin d'être uniforme. Ce contexte représente un premier défi pour une surveillance adéquate des maladies chroniques. Ainsi, le fichier des décès est administré par le MSSS conjointement avec l'Institut de la statistique du Québec (ISQ). Son rôle principal est de contribuer au suivi démographique de la population québécoise. Il s'agit d'un fichier validé et revalidé, dont la qualité des données est excellente. Du côté des fichiers des hospitalisations, il est géré par le MSSS et son utilité première est le suivi de la morbidité hospitalière et l'utilisation des ressources de santé. Encore une fois, la qualité des données n'est pas remise en question, et de nombreuses causes d'hospitalisations permettent d'identifier plusieurs maladies. Dans le passé, avant l'ère du jumelage, les fichiers de décès et d'hospitalisations étaient au cœur de la surveillance des maladies chroniques. Ce sont donc des entités connues.

Les trois autres nouveaux fichiers proviennent de la Régie de l'assurance maladie du Québec (RAMQ). Bien qu'il soit administré par le même organisme, leur utilité principale est bien différente. Le fichier des personnes assurées permet tout simplement d'effectuer le suivi des personnes assurées, et donc de déterminer quels citoyens sont assurés par le régime d'assurance public, et à quel moment dans le temps. Le FIPA contient aussi des informations sociodémographiques sur les individus, notamment le lieu de résidence, l'âge et le sexe. L'objectif central du fichier des services médicaux est le paiement des médecins en fonction des actes médicaux réalisés. De sorte que la variable du code d'acte médical est la plus importante, et par conséquent la mieux validée. Or, pour identifier la plupart des maladies chroniques, c'est la variable du code de diagnostic qui est plus importante. La validation de cette dernière n'est pas aussi minutieuse. En pratique, le code de diagnostic n'est pas toujours présent. Finalement, le fichier des services pharmaceutiques sert principalement à rembourser les assurées du régime public d'assurance médicaments. Bien que certains médicaments soient clairement associés à une maladie chronique, le fichier des services pharmaceutiques n'identifie pas directement de quelle(s) maladie(s) souffrent les personnes assurées puisque ce n'est pas son rôle premier.

Des différences entre les cinq fichiers viennent aussi compliquer l'identification de maladies chroniques, ou encore la comparabilité des estimations dans le temps. En premier lieu, le passage de la neuvième classification internationale des maladies (CIM-9) vers la dixième version (CIM-10) n'a pas été effectué au même moment. Au Québec, ce passage a été réalisé en 2000 pour le fichier des décès, en 2006 pour le fichier des hospitalisations et il n'a pas encore été effectué pour le fichier des services médicaux. Les codes de diagnostic de ce dernier sont encore codés selon CIM-9. Déterminer l'impact de cette discordance sur les estimations est un défi qui a dû être considéré et étudié.

À cela s'ajoute le fait que le fichier des décès accuse un certain retard par rapport aux autres fichiers. Tel qu'indiqué ci-haut, ce fichier est validé minutieusement, ce qui rallonge le processus de diffusion auprès des utilisateurs. Surtout, certains décès (la plupart accidentels) requièrent l'implication du Bureau du coroner. Les enquêtes du coroner peuvent nécessiter plusieurs mois voire plus d'un an, ce qui retarde la diffusion du fichier des décès. Par conséquent, tout indicateur de surveillance maladies chroniques impliquant le fichier des décès couvre une période moins longue et moins récente. À noter que le FIPA contient aussi une date de décès, mais il n'inclut aucune cause de décès.

La dernière principale différence entre les fichiers est la population couverte par chacun d'entre eux. Alors que les fichiers de décès et d'hospitalisations couvrent l'ensemble de la population, sans exception, le FIPA et le fichier des services médicaux touchent seulement les Québécois assurés au régime d'assurance maladie. De sorte que les informations sur certains groupes de la population, notamment les militaires, les personnes vivant hors Québec pendant la majorité d'une année et les personnes ayant recours au système privé de santé, ne sont pas disponibles. Enfin, au moment d'écrire ces lignes, le fichier des services pharmaceutiques couvre seulement les Québécois de 65 ans et plus qui sont assurés par le régime public, ce qui exclut non seulement les individus couverts par une assurance privé, mais aussi ceux vivant en institution qui reçoivent leurs médicaments directement via cette

institution. La couverture variable de la population impose des défis, au premier chef pour la production d'estimations populationnelles à partir de bases de données qui ne sont pas tout à fait populationnelles.

2. Qualité des données

2.1 Qualité inégale entre fichiers et entre variables

Tel que spécifié précédemment, les fichiers de décès et d'hospitalisations sont méticuleusement validés, et ce pour l'ensemble des variables présentes. La qualité est beaucoup plus variable pour les trois fichiers de la RAMQ, non seulement entre fichiers, mais aussi à l'intérieur de chaque fichier. D'une variable à l'autre, la fiabilité des informations n'est pas constante, ce qui peut évidemment d'avoir des impacts sur les estimations produites. Reprenons l'exemple du fichier des services médicaux, dans lequel le code d'acte est d'une excellente qualité, puisque le paiement des médecins repose en grande partie sur cette variable. Autrement dit, autant la RAMQ et les médecins ont intérêt à ce que les actes médicaux soient codifiés adéquatement et avec précision. Ce qui n'est pas nécessairement le cas pour le code de diagnostic qui accompagne le code d'acte, puisque le diagnostic n'a aucun impact direct sur le paiement des médecins (sauf exception). Dans l'exemple ci-dessous, il est possible d'identifier la maladie chronique qui affecte le premier patient, mais pas pour le second patient étant donné l'absence du code de diagnostic.

Tableau 2.1.1
Exemple de qualité inégale entre variables d'un même fichier

Code d'acte	Code de diagnostic (CIM-9)	Maladie
09162 – Visite principale	250	Diabète
09162 – Visite principale	?	?

Cet exemple montre l'impact direct que peut avoir une seule variable sur la qualité des estimations produites. La section suivante mettra en relief d'autres situations où la qualité des variables est problématique, mais où l'impact est parfois indirect. Certaines incohérences entre les fichiers seront aussi expliquées.

2.2 Problèmes de qualité : exemples et incohérences

Une des variables les plus importantes du SISMACQ est le code postal des individus que l'on retrouve dans le fichier des personnes assurées. Le code postal permet notamment d'identifier le lieu de résidence de chacun et de suivre leurs déplacements au fil des ans. Les estimations par régions sociosanitaires et par territoires plus fins sont étroitement dépendantes du code postal. Or, la qualité de cette variable laisse à désirer. D'abord, au début de la période (particulièrement en 1996), de nombreux codes postaux sont tout simplement manquants. Ensuite, tout au long de la période, une proportion faible mais non négligeable de codes postaux sont erronés, la plupart du temps le résultat d'erreurs de saisie. Finalement, à l'occasion, certains « trous » dans les adresses ne nous permettent pas de connaître le lieu de résidence d'un individu pendant une certaine période de temps plus ou moins longue. Plus concrètement, la date de fin d'une adresse (code postal) ne correspond pas à la date de début de l'adresse subséquente.

Le code municipal, aussi appelée le code de la subdivision de recensement, est aussi imparfait. La RAMQ assigne ce code municipal à partir du code postal, Or, en milieu rural, plusieurs municipalités peuvent avoir le même code postal. Dans ces situations, le processus d'assignation utilisé par la RAMQ fait en sorte que le code municipal de la municipalité la plus populeuse est assigné. De sorte que, selon le fichier du FIPA reçu par l'INSPQ, environ 250 municipalités n'ont aucun habitant. L'INSPQ a donc développé, avec l'aide du nom de la municipalité qui est aussi une variable présente dans le fichier, sa propre assignation du code municipal afin combler cette lacune. Cependant, le nom de la municipalité est malheureusement aussi une variable dont la qualité laisse à désirer. En effet, la RAMQ n'uniformise pas les noms, de sorte qu'à ce jour les habitants de Montréal ont 1183 noms de municipalités différents, et les habitants de Québec ont 674 noms différents. Les variations incluent de simples d'erreurs d'orthographes (ex : Monréal), des noms de municipalités pré-fusion (ex : Dorval), des ajouts de préfixes ou de suffixes (ex : Montréal QC), des adresses quasi complètes (ex : rue Sherbrooke, Montréal) ou des combinaisons de ces facteurs. Afin d'utiliser ce nom de municipalité, l'INSPQ doit procéder à une certaine uniformisation lorsque possible.

Certaines incohérences affectent aussi la qualité des données, et par extension la qualité des estimations produites. Dans un premier exemple, la date de décès qu'on retrouve dans le fichier des personnes assurées n'est pas toujours la même que celle dans le fichier de décès. Lorsque possible, c'est cette dernière qui doit primer étant donné la validation nettement supérieure du fichier des décès. Pour la RAMQ, la date de décès n'a pas à être très précise, l'important étant de savoir qu'un individu est décédé et donc qu'il n'est plus admissible au régime d'assurance publique. En deuxième lieu, les services pharmaceutiques apparaissent parfois pour des individus qui ne sont pourtant pas admissibles selon la variable d'admissibilité au régime d'assurance public. Des délais administratifs expliquent généralement ce type d'incohérence. Cela signifie que la variable d'admissibilité n'est pas entièrement fiable. Troisièmement, la population des 20 à 30 ans est passablement sous-estimée par le FIPA, pour la simple et bonne raison que les jeunes adultes renouvellent moins leur carte d'assurance-maladie. Ils n'ont plus leurs parents pour le faire à leur place, et ils sont généralement en santé et donc ils ne consomment pas de soins de santé régulièrement. Finalement, la mise à jour annuelle des données du SISMACQ, qui se déroule au printemps de chaque année, peut modifier légèrement les estimations obtenues avec la version précédente du système, même pour des années antérieures éloignées. Pour différentes raisons, la RAMQ applique des changements à ces données (ajouts, retraits, modifications ou corrections). À noter que les fichiers de décès et d'hospitalisations ne changent pas. Seule une année de données est ajoutée.

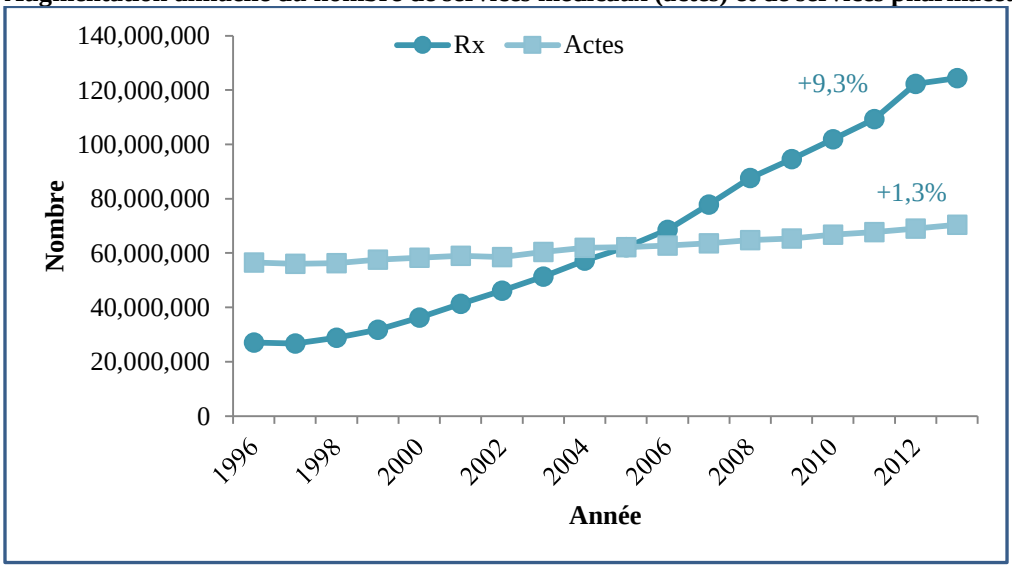
3. Exploitation simultanée des données

3.1 Taille

Les données administratives sont en quelque sorte les premières données massives. Elles ne sont pas le résultat d'une enquête planifiée à l'avance par un ou plusieurs chercheurs, mais plutôt de l'activité humaine quotidienne. Dans le cas qui nous concerne, chaque événement sanitaire (décès, hospitalisation, consultation, achat de médicaments) fait l'objet d'un enregistrement, pour la très grande majorité des huit millions de Québécois, et ce sur une période de près de 20 ans (1996 à 2014). La taille des données ne fait qu'augmenter dans le temps, ce qui entraîne certains défis lors de l'utilisation et l'analyse.

Sur la période, on compte près de 900 000 décès, 19,5 millions d'hospitalisations, 1,13 milliard de services médicaux et 1,22 milliard de services pharmaceutiques, représentant chacun une observation détaillée sur plusieurs variables. Ces données sont stockées sur un serveur local et sont accédés via le logiciel SAS Enterprise Guide. Pour les décès et les hospitalisations, le temps de lecture des bases de données complètes, sans effectuer de modification, est court. Il est respectivement de six et de vingt secondes. Cependant, pour les services médicaux et les services pharmaceutiques, le temps d'attente des fichiers complets est aujourd'hui de 75 et de 120 minutes, respectivement. Les manipulations subséquentes de ces fichiers requièrent aussi beaucoup de temps, et le graphique ci-dessous démontrent que cette situation va perdurer malgré les avancées technologiques et informatiques.

Tableau 3.1.1
Augmentation annuelle du nombre de services médicaux (actes) et de services pharmaceutiques (Rx)



Le temps d'attente n'est pas la seule conséquence de la quantité d'information à traiter. Les données doivent aussi être stockées, et l'espace de stockage doit être gérer convenablement dans un contexte où le nombre d'utilisateurs est relativement élevé.

3.2 Stratégies pour éviter le temps perdu et les problèmes de stockage

Pour travailler efficacement, il est impensable d'attendre des heures à chaque jour simplement pour traiter les fichiers originaux. Il est donc crucial de créer des bases de données permanentes transformées. Ces dernières conservent seulement les observations nécessaires à la réalisation d'un ou plusieurs projets. Ensuite, pour les tâches plus lourdes, utiliser les temps morts, c'est-à-dire les midis, les soirées et les fins de semaine, est la clé de pour être plus efficace. Particulièrement en considérant que plusieurs personnes sont connectées au serveur local pour une même journée donnée. Se servir des temps morts permet aussi d'éviter de ralentir les autres utilisateurs en exécutant une requête massive.
La création de données permanentes transformées, qui sont souvent transitoires, peut toutefois entraîner un effet indésirable, soit la diminution rapide de l'espace de stockage disponible. Il devient alors important d'optimiser l'espace de stockage permanent et l'espace de stockage temporaire (qui n'est pas conservée à la fermeture du programme). Dans un monde idéal, le « cloud computing », par sa scalabilité et son élasticité, permettrait de résoudre ce défi de manière élégante, mais la section 4 de ce document expliquera pourquoi le SISMACQ ne peut bénéficier de cette technologie pour le moment. La meilleure alternative, qui est d'augmenter la taille de l'espace de stockage sur le serveur local, n'est pas toujours une solution envisageable, particulièrement dans un contexte de restrictions budgétaires.

En attendant, il existe quelques autres solutions pour sauver du temps et améliorer l'efficacité des utilisateurs de la base de données. Une d'entre elles est de prévoir un ou plusieurs ordinateurs supplémentaires qui sont dédiées à des requêtes ou à des analyses lourdes. De sorte que, pendant qu'une de ces requêtes lourdes est complétée par un ordinateur additionnel, l'utilisateur peut continuer à travailler sur son poste personnel sur un autre projet ou un autre volet du même projet. Aussi, pour obtenir des résultats préliminaires, travailler sur un échantillon de la population permet d'accélérer le processus. Généralement, les résultats obtenus à partir d'un tel échantillon demeurent représentatifs étant donné la taille du SISMACQ. Par exemple, sélectionner 10% des individus du SISMACQ permet tout de même de travailler sur près d'un million d'individus, une cohorte amplement suffisante pour la très grande majorité des analyses. Cette façon de faire évite de devoir exécuter des modèles statistiques préparatoires nécessitant plusieurs heures d'attente, tels que les modèles de survie et les analyses multiniveaux ou hiérarchiques.

Enfin, deux éléments plus techniques qui peuvent éviter bien des tracas. D'abord, la planification adéquate des périodes de maintenance avec l'équipe des technologies de l'information est primordiale afin d'éviter les mauvaises surprises. Des périodes de maintenance à intervalle régulier facilitent grandement la planification pour les utilisateurs de données. Aussi, la fiabilité du réseau électrique doit être excellente afin d'éviter des pannes de courant au mauvais moment. Cette situation est survenue à quelques reprises à l'INSPQ, ses bureaux de Québec étant malheureusement situé dans un secteur où le réseau électrique est particulièrement sensible aux aléas de la nature.

4. Autres contraintes

4.1 Contraintes légales

Le SISMACQ a été créé suite à une entente tripartite entre l'INSPQ, le MSSS et la RAMQ, en plus d'être approuvé par la Commission d'accès à l'information. Chaque modification à l'entente, même mineure, doit être approuvée par tous les organismes impliqués. De sorte que l'ajout d'information est un processus long, ardu, et qui fait souvent l'objet de délais. Que ce soit pour l'ajout d'une seul variable comme le code d'accident permettant d'identifier un traumatisme, ou encore l'inclusion de tous les services pharmaceutiques, même chez les moins de 65 ans. Aussi, le SISMACQ peut seulement être utilisé à des fins de surveillance des maladies chroniques. Il n'est donc pas permis de s'en servir pour étudier les maladies infectieuses ou les traumatismes, pour ne donner que deux exemples.

L'entente encadre aussi toutes les étapes de la manipulation des données, principalement afin de respecter la sécurité de l'information et la confidentialité des Québécois. Ainsi, le transport des données de la RAMQ vers l'INSPQ est balisé par certaines règles, tout comme l'entreposage et le stockage de ces mêmes données. C'est ce

qui explique pourquoi il est présentement impossible de considérer le « cloud computing » comme solution de stockage. En effet, l'entente stipule actuellement que les données doivent se retrouver sur un serveur local sécurisé afin de maximiser la sécurité de l'information. Pour la même raison, l'accès physique aux locaux est restreint aux employés chargés de la surveillance des maladies chroniques, et l'accès aux postes informatiques est encadré selon différents statuts d'utilisateurs. Par exemple, seuls quelques utilisateurs ont accès à des variables sensibles comme le code postal, la date de naissance et la date de décès. Ces accès informatiques sont journalisés dans l'éventualité où un problème surviendrait. Finalement, la diffusion des résultats doit aussi respectée plusieurs règles de sécurité, toujours afin de minimiser les risques de bris de confidentialité.

4.2 Autres sources de données

Le SISMACQ permet l'ajout de variables contextuelles et territoriales importantes, le plus souvent à partir du code postal. C'est ainsi que des territoires comme la région sociosanitaire est assignée à tous les individus ayant un code postal valide, afin de produire des estimations régionales ou locales. Une des faiblesses des bases de données administratives est l'absence d'informations à caractère socio-économiques. Toujours via le code postal, il est possible d'ajouter ce type d'information via des proxys écologiques et l'indice de défavorisation. Cependant, l'ajout d'informations individuelles n'est pas permis. C'est ainsi que le jumelage du SISMACQ avec des données d'enquête est interdite pour l'instant. Ce type de jumelage permettrait notamment de considérer des habitudes vie comme le tabac et la nutrition dans nos analyses, les habitudes de vie étant étroitement associées à plusieurs maladies chroniques. Le fait de ne pouvoir les considérer est un défi important.

Au Canada, certaines provinces peuvent déjà compter sur des dossiers électroniques médicaux comme source de données supplémentaire. Au Québec, ce n'est pas encore le cas. Il ne fait pas de doute que d'éventuels dossiers électroniques médicaux, s'ils deviennent accessibles, permettront d'améliorer davantage la surveillance des maladies chroniques. Il n'en demeure pas moins que le SISMACQ, tel qu'actuellement constitué, est une excellente source de données permettant de calculer de bien meilleures estimations que par le passé. Son coût est aussi relativement modeste, compte tenu du fait que les données existent déjà. En somme, les défis détaillés dans ce manuscrit sont importants, mais les avantages sont si nombreux que personne ne désire revenir en arrière.

Bibliographie

Blais C, Jean S, Sirois C, Rochette L, Plante C, Larocque I, Doucet M, Ruel G, Simard M, Gamache P, Hamel D, St-Laurent D, Émond V (2014), « Le système intégré de surveillance des maladies chroniques du Québec (SISMACQ), une approche novatrice (Quebec Integrated Chronic disease surveillance system (QICDSS), an Innovative approach)», *Maladies chroniques et blessures au Canada*, Vol. 34 no 4, p. 247-256.

Geran L, Tully P, Wood P, Thomas B. (2005) Comparability of ICD-10 and ICD-9 for Mortality Statistics in Canada, *Statistics Canada*.

Pampalon R, Hamel D, Gamache P, Raymond G. (2009) A deprivation index for health planning in Canada. *Chronic Dis Can*; 29(4):178-191.

Régie de l'assurance maladie. (2016) Manuel de facturation – Rémunération à l'acte, Mise à jour 89, Janvier 2015, 874 pages.

Rochette L, Émond V. (2004) Surveillance des maladies chroniques au Québec par le jumelage des fichiers administratifs. *Recueil du Symposium 2014 de Statistique Canada*.