Recueil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

ENQUÊTES SUR LES DONNÉES : MISE EN ÉVIDENCE D'ERREURS SYSTÉMATIQUES DANS LES BASES DE DONNÉES ADMINISTRATIVES

Sten Ardal[1] et Sherri Ennis[1]

RÉSUMÉ

Les utilisateurs secondaires des données sur la santé supposent souvent que les données administratives représentent une base assez solide pour servir de fondement à d'importantes décisions stratégiques et de planification. Si elles sont distribuées uniformément et(ou) aléatoirement, les erreurs qui entachent les données ont généralement peu d'effet sur les conclusions d'une analyse. Toutefois, si les sources de données contiennent des erreurs systématiques ou si de telles erreurs sont introduites au moment de la création des fichiers maître, les conclusions de l'analyse risquent d'être erronées. Les erreurs systématiques les plus courantes comprennent la sous-estimation de l'activité d'une population donnée, le recodage inexact des renseignements spatiaux ou le manque d'uniformité des protocoles de saisie des données. Grâce à un partenariat entre les régions de planification sanitaire les plus peuplées de l'Ontario, le Central East Health Information Partnership (CEHIP) fournit l'information requise pour élaborer les programmes de santé publique et planifier le système de santé. Le CEHIP a repéré un certain nombre d'erreurs systématiques dans les bases de données administratives et en a décrit un grand nombre dans des rapports qui ont été envoyés aux organismes partenaires. Il a étudié les cas de naissances non enregistrées, ainsi que l'attribution incorrecte des codes géographiques dans les fichiers de l'état civil. Il a aussi examiné les problèmes de codage de la cause de décès, particulièrement en ce qui a trait au retard avec lequel elle est établie et à l'effet de ce retard sur les ensembles de données officielles. Enfin, l'étude des différences entre les protocoles de saisie des données sur les maladies à déclaration obligatoire soulève des questions quant à la cohérence des données transmises par les divers organismes de surveillance. Le présent article expose comment certaines de ces erreurs ont été repérées et précise quels processus donnent lieu à cette perte d'intégrité des données. En conclusion, il décrit certaines conséquences de ces problèmes pour les planificateurs des programmes de santé, les gestionnaires des programmes et les décideurs.

MOTS CLÉS : Qualité des données; information sur la santé; planification sanitaire.

1. INTRODUCTION

> *« Winwood Reade est intéressant sur ce sujet. Il remarque que, tandis que l'individu pris isolément est un puzzle insoluble, il devient, au sein d'une masse, une certitude mathématique. Par exemple, vous ne pouvez jamais prédire ce que fera tel ou tel, mais vous pouvez prévoir comment se comportera un groupe. Les individus varient, mais la moyenne reste constante. Ainsi parle le statisticien. » (*Sherlock Holmes, *Le Signe des Quatre*)

En matière de santé de la population, la planification des services et l'élaboration des politiques reposent sur la « certitude » que donne l'analyse de données sur de grands groupes d'individus. Pourtant, les renseignements sur lesquels nous nous fondons pour décrire les caractéristiques d'une population et mesurer ses actions ne produisent pas toujours une moyenne « exacte ». La moyenne observée dépend de l'erreur de mesure et est décalée en cas d'erreur systématique.

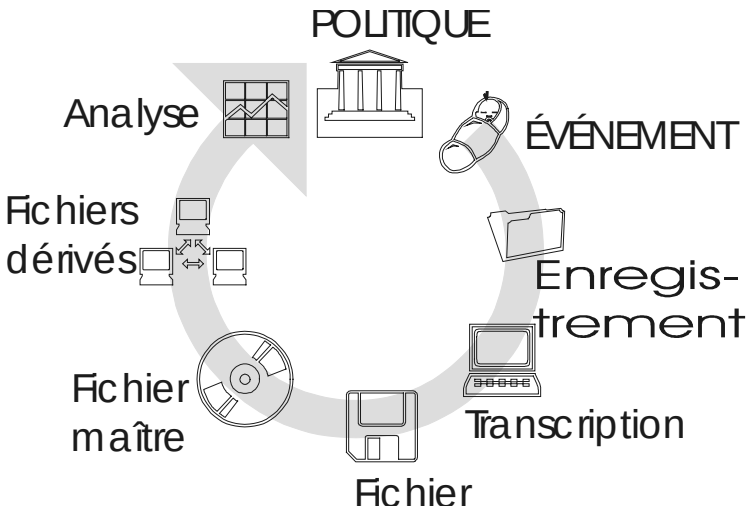
[1] Central East Health Information Partnership
C.P. 234 Newmarket (Ontario), L3Y 4X1
Téléphone : 905.764.6346 poste1211, télécopieur : 905.895.0848
Courriel : info@cehip.org

En statistique, on suppose habituellement que l'« erreur » obéit à une loi de distribution aléatoire qui produit la courbe normale utilisée pour les tests de probabilité. L'erreur est imputée à la précision de l'instrument de mesure utilisé pour l'étude (Fisher,1949). Bien qu'on puisse l'appliquer à l'interprétation des données d'enquête, cette notion d'erreur diffère considérablement du genre d'erreurs observées lorsque l'on analyse des données administratives. La plupart des données utilisées pour décrire la santé de la population sont recueillies dans le cadre de la prestation de services. Les erreurs produites durant la création d'ensembles de données administratives reflètent les méthodes de collecte et de traitement des données et, le plus souvent, sont de nature systématique (Wolff et Helminiak, 1996). Nous montrons dans le présent article comment le Central East Health Information Partnership (CEHIP) a décelé certaines erreurs dans des ensembles de données utilisés couramment et indiquons les procédures qui ont biaisé les données figurant dans certaines zones importantes des enregistrements. À cette fin, nous décrivons le déroulement du processus de traitement des données. Enfin, nous décrivons la scène du crime, donnons les détails de notre enquête et identifions les coupables.

> *« Et bâtir une théorie avant d'avoir des données est une erreur monumentale : insensiblement on se met à torturer les faits pour qu'ils collent avec la théorie, alors que ce sont les théories qui doivent coller avec les faits. » (*Sherlock Holmes*, Un scandale en Bohème)*

Il est tentant de vouloir accepter les résultats d'une analyse, particulièrement s'ils concordent avec ceux attendus. Nous considérons que les données sont « exactes » si elles correspondent à notre cadre théorique, mais blâmons la qualité de l'information si elles paraissent aberrantes. Pourtant, les données que manipule l'analyste reflètent toujours les processus et les méthodes utilisés pour les produire. Avant de nous lancer dans leur interprétation, nous devrions nous assurer qu'aucun fait important n'a été oublié. Nous devrions accorder une attention particulière aux aspects de la collecte et du traitement des données susceptibles de causer des erreurs.

Donc, quels sont les faits dont nous devons tenir compte? La figure 1 donne le schéma général du cheminement des données administratives.



Événement : Peut ne pas être déclaré.

Enregistrement : Mauvais renseignements.

Transcription : Manque d'uniformité des protocoles de saisie utilisés.

Fichier : Structure ne permettant pas de décrire convenablement les événements.

Fichier maître : Critères d'exclusion.

Fichiers dérivés : Erreurs d'enregistrement.

Analyses : Les totalisations peuvent produire le cumul des erreurs.

Politique : Peut introduire un biais systématique de déclaration.

Figure 1 : *Diagramme simplifié de cheminement de l'information avec exemples de divers facteurs d'erreur à des étapes particulières du traitement de l'information.*

Ce cheminement peut comprendre une foule de processus complexes susceptibles d'être influencés par les événements qui surviennent aux étapes illustrées à la figure 1, ainsi qu'entre ces étapes. Entre les étapes, la non-saisie ou la non-transmission des données est l'événement le plus probable. Par contre, à chaque étape, un grand nombre de processus peuvent éventuellement entrer en jeu. Un groupe ontarien, appelé Clinical Data Quality Task Force, s'est efforcé de comprendre le fonctionnement de certains de ces processus dans les établissements de soins. La figure 2 donne une idée des personnes, des structures et des processus qui contribuent à la création d'un enregistrement type. C'est ce modèle que nous utilisons pour étudier les effets des modifications de la structure organisationnelle sur la qualité des données. Par exemple, la tendance à décentraliser les procédures d'enregistrement des patients pourrait nécessiter diverses formes de formation, de surveillance et d'interfaces en vue de maintenir le niveau voulu de qualité.

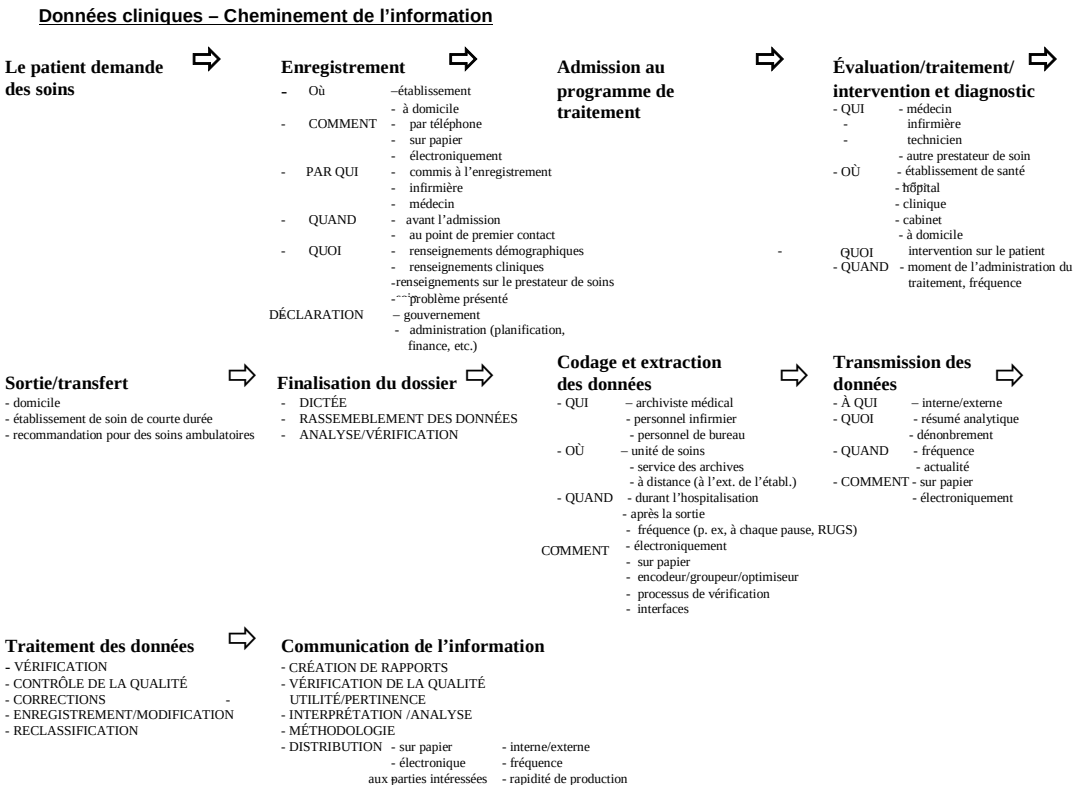


Figure 2 : *Diagramme de cheminement des données cliniques pour un grand établissement de soins type*

Manifestement, le chemin parcouru par les renseignements administratifs est complexe. Le CEHIP a constaté que diverses bases de données administratives posent des problèmes. Les formes d'erreurs qui peuvent être introduites dans un système complexe sont fort nombreuses. Comme il est probable que seules les plus flagrantes soient effectivement remarquées, le dépistage des erreurs systématiques que l'on peut éliminer grâce à l'amélioration des processus contribuera à une meilleure interprétation des données, mais n'aboutira jamais à une exactitude totale. Nous présentons trois « études de cas » pour illustrer les enquêtes sur la qualité des données menées par le CEHIP.

2. EXEMPLES D'ÉTUDES DE LA QUALITÉ DES DONNÉES

2.1 Le cas des bébés disparus

La scène : Le CEHIP a remarqué que le nombre de naissances enregistrées dans les dossiers statistiques des hôpitaux diffèrent du nombre déclaré par le Bureau du registraire général de l'Ontario. Les dénombrements faits par les hôpitaux sont généralement plus élevés, alors que, logiquement, ils devraient être légèrement inférieurs, puisque quelques naissances ont lieu dans des hôpitaux à l'extérieur de l'Ontario. Lors de discussions avec des représentants du ministère provincial de la Santé et de Statistique Canada, nous nous sommes aperçu que les deux organismes avaient, eux aussi, des réserves quant à la complétude des renseignements qui figurent dans le registre des naissances. On savait qu'une politique adoptée il y a quelques années exigeait qu'un avis de naissance soit présenté par un parent ainsi que par le médecin pour que le fichier maître soit mis à jour, mais on ne disposait d'aucun renseignement sur le nombre d'avis qui n'avaient pas donné lieu à une entrée dans un fichier de l'état civil.

L'enquête : À notre avis, le meilleur moyen de lancer l'enquête consistait à inclure les entrées qui ne répondaient pas au critère de l'« avis double ». Cette décision a nécessité la saisie manuelle des données de milliers de dossiers. Dans presque tous les cas, l'avis manquant était celui qui doit être présenté par le parent. Aux fins de l'analyse, nous avons créé un nouveau fichier maître qui comprenait tous les rapports de naissance reçus par le registraire général. L'analyse de ce fichier a révélé que les jeunes parents dont le bébé avait un faible poids à la naissance étaient ceux qui étaient les plus susceptibles de ne pas déclarer cette dernière et, de surcroît, que la situation variait considérablement selon la région. Il convient de souligner que l'analyse n'incluait pas les événements pour lesquels la probabilité d'une naissance vivante était faible. Nous avons constaté que le taux de non-déclaration avait augmenté autour de 1996.

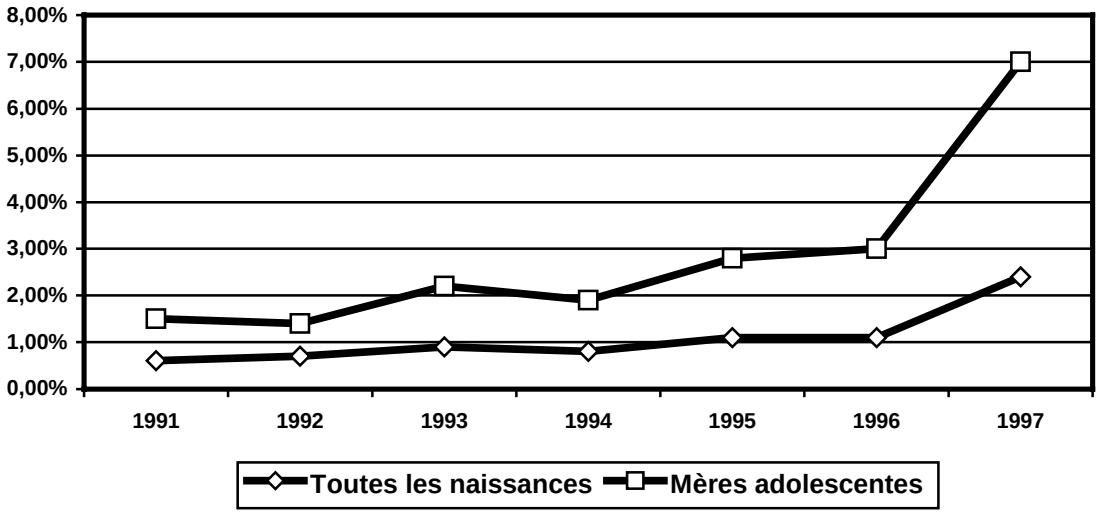


Figure 3 : *Pourcentage de naissances non déclarées repérées en Ontario de 1991 à 1997 pour l'ensemble des naissances et pour les naissances imputables à des adolescentes.*

Que s'est-il passé en 1996? Cette année-là, la province a mis en application une loi permettant le recouvrement du coût des enregistrements de naissance faits par les parents, service qui auparavant était offert gratuitement au nom du registraire provincial. Nous avons donc mené une enquête auprès de toutes les municipalités ontariennes afin de déterminer si elles percevaient des frais et, le cas échéant, quel en était le montant et à quel moment elles avaient commencé à le faire. Après l'ajout de ces renseignements dans l'analyse, il est devenu évident que l'augmentation la plus importante de « naissances manquantes » avait eu lieu dans les municipalités qui avaient commencé à percevoir des frais. Ces dernières représentaient environ 70 % de l'ensemble des naissances survenues en Ontario.

Les résultats : Cet examen de la qualité des données montre que l'introduction d'une erreur systématique a causé la sous-représentation d'une sous-population particulière. Comme il s'agit d'un groupe visé par des campagnes spéciales de promotion de la santé, ce sous-dénombrement peut avoir un effet sur la configuration des services, l'analyse des tendances clés et la politique publique. Si les données n'étaient pas corrigées, les taux de bébés de faible poids à la naissance, de grossesses chez les adolescentes et de croissance de la population seraient sous-estimés. Les résultats de l'étude ont été communiqués au Bureau du registraire général de l'Ontario et devraient, en principe, influencer les politiques d'enregistrement et de compilation des données.

2.2 Le mystère de la fécondité fluctuante

La scène : Les taux de fécondité sont des indicateurs importants utilisés pour planifier un large éventail de services à la population. L'évolution des tendances influence considérablement les décisions prises à l'échelon local, comme la construction d'écoles, et les politiques nationales, comme celles régissant les taux d'immigration. L'analyse de routine d'un fichier du registre des naissances de l'Ontario transmis à Statistique Canada a indiqué une diminution soudaine, de l'ordre de plusieurs centaines, du nombre de naissances dans certaines régions et une augmentation dans d'autres. Le calcul des taux de fécondité a montré que ces variations n'étaient pas imputables à des modifications de la répartition démographique des mères potentielles.

L'enquête : Il a été assez facile d'établir qu'une faute visant les données avait été commise. Dans la région de Markham, le nombre de naissances avait augmenté d'environ 600, soit 40 %. Entre-temps, dans la région de Vaughan, il avait diminué de moitié, soit d'environ 850. De surcroît, nous avons constaté que les dénombrements de naissances faits par les régions étudiées ne concordaient pas avec ceux faits par les hôpitaux. Or, comme virtuellement toutes les naissances ont lieu à l'hôpital, on s'attendrait à ce que les chiffres concordent. L'ampleur du problème est devenu évidente lorsque nous avons calculé les taux de fécondité afin d'obtenir des indicateurs comparables tenant compte de la taille de la population en âge de procréer. Les taux ont révélé des tendances divergentes pour les régions étudiées lorsque la ventilation a été faite en fonction des codes de la subdivision de recensement (SDR) attribués par Statistique Canada d'après les renseignements municipaux contenus dans les fichiers transmis par le Bureau du registraire général de l'Ontario.

Les fichiers reçus contenaient aussi les codes postaux. Le recodage des SDR d'après les codes postaux a produit des tendances plus stables, concordant mieux avec les résultats de l'analyse des dossiers hospitaliers. En outre, les données recodées sur les subdivisions de recensement et les renseignements municipaux figurant dans le fichier concordaient.

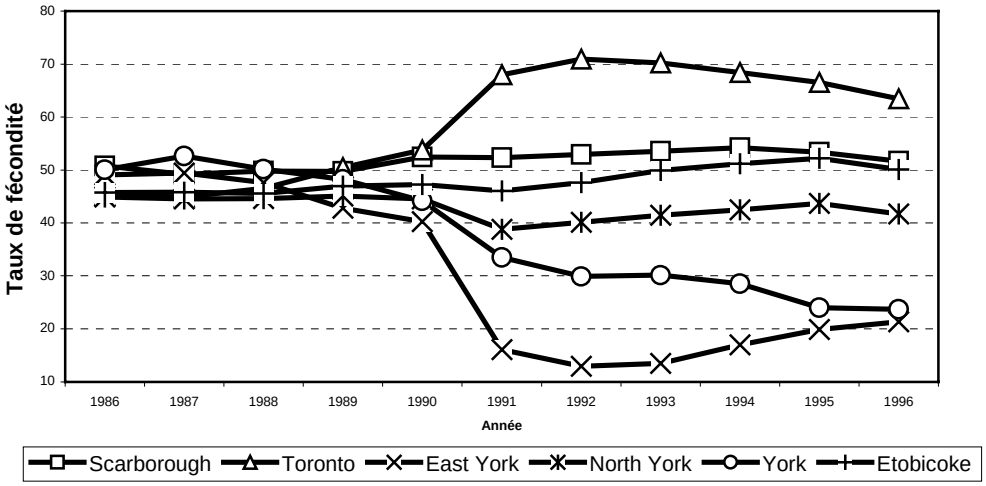


Figure 4 : *Taux de fécondité pour certaines municipalités de l'Ontario, 1986 à 1995.*

Les résultats : Non corrigée, cette erreur donnerait lieu à des inférences incorrectes quant à la croissance de la population et au besoin de services à la famille et à l'enfance. Après avoir repéré l'erreur et averti les personnes chargées de créer le fichier, nous avons rédigé une note sur la discordance observée en vue d'informer les responsables de la planification et de l'élaboration des politiques qui se servent de ces données. Nous n'avons pas élucidé entièrement le problème, mais nous supposons que les personnes qui ont rempli les déclarations ont été moins précises en ce qui concerne leur municipalité que leur code postal. Dans la région de Toronto, une région métropolitaine contiguë était subdivisée en plusieurs « municipalités » qui ont été fusionnées depuis. De toute évidence, les processus de vérification étaient inadéquats à toutes les étapes du traitement des données. En effet, celles des municipalités auraient pu être comparées aux adresses de voirie ou aux codes postaux durant l'enregistrement. En outre, la cohérence interne des données géographiques dans les fichiers aurait pu être confirmée avant l'attribution des codes de SDR.

2.3 Le cas du calendrier

La scène : Avant 1998, le Grand Toronto était divisé en six municipalités, dotées chacune de son propre département de santé publique. La fusion qui a eu lieu en 1999 a soulevé des questions quant à la comparabilité des anciens systèmes informatiques sur les maladies à déclaration obligatoire (SIMDO) des municipalités. Ces systèmes sont censés fournir des renseignements comparables et cohérents qui sont regroupés dans la base de données provinciale. Le dégagement des tendances importantes et la surveillance des flambées dépendent de l'obtention de données d'identification et de suivi correctes sur les 60 maladies que les bureaux de santé sont tenus par la loi de déclarer au gouvernement de l'Ontario. Nous soupçonnions l'existence de discordance entre les méthodes de codage des données des anciennes municipalités et avons décidé de mener une enquête afin de repérer les problèmes de qualité liés à la fusion des bases de données des six SIMDO de Toronto.

L'enquête : Nous avons interrogé des intervenants clés afin de comprendre le processus utilisé pour enregistrer l'information. Cela nous a permis de constater qu'elle était souvent saisie par des commis connaissant à peine les protocoles de saisie des données et que, très souvent, un grand nombre de personnes participaient aux opérations de saisie.

Nous avons examiné toutes les zones des enregistrements, que leur remplissage soit obligatoire ou facultatif, afin de déterminer si l'information enregistrée était incomplète, incohérente ou, en apparence, inexacte. Nous nous sommes concentrés sur quatre des maladies déclarées le plus couramment. Le taux d'erreurs variait d'une valeur virtuellement nulle pour certaines zones à saisie obligatoire à presque 100 % pour certaines zones à saisie facultative.

Les zones de date constituent un bon exemple d'incohérence, puisqu'il est assez facile de repérer les dates logiquement impossibles. Cela a été le cas lors de notre analyse des fichiers de données sur la grippe, dont 6,4 % des enregistrements contenaient une date de manifestation de la maladie ultérieure à celle du diagnostic médical et 7,6 % contenaient une date de déclaration de la maladie au bureau de santé antérieure à la date du diagnostic. Le système est conçu pour enregistrer un cas soupçonné, puis la date subséquente de diagnostic et enfin la date subséquente du rapport diagnostique (voir le tableau 1). En analyse épidémiologique des flambées de maladie, l'exactitude des dates enregistrées peut être critique. Or, le taux réel d'erreur est vraisemblablement plus élevé que le taux observé de données illogiques.

Séquence logique	*Date d'enregistrement du cas soupçonné*	*Date du diagnostic*	*Date du rapport diagnostique*
Pourcentage de données « suspectes »	7,5 %	14,4 %	8,8 %

Tableau 1 : *Pourcentage de zones de date, pour les cas de grippe déclarés au Système informatique sur les maladies à déclaration obligatoire de Toronto, qui sont précédées ou suivies par l'enregistrement d'une date associée logiquement impossible.*

Les résultats : Non corrigées, ces erreurs fausseraient les analyses de tendance et retarderaient l'utilisation des données, alors qu'elles sont destinées à repérer rapidement les épidémies. Les procédures de saisie utilisées se fondent sur des définitions fort variables des zones de données. Par exemple, la « date de l'épisode » était interprétée variablement comme signifiant « apparition des symptômes » (correct), « date de la visite chez le clinicien », « date du prélèvement d'échantillon », « date de la saisie des données » ou « date de la finalisation du rapport ». Pareillement, la « date du diagnostic » était interprétée comme signifiant la date de réception du rapport par le bureau de santé, la date du diagnostic (correct) ou la date déclarée par un client.

La qualité de ces ensembles de données souffrait du manque de respect des protocoles de saisie des données, dû fort probablement à une formation inadéquate et à l'absence de procédures de contrôle de la qualité. Ce genre d'erreur a été bien décrit (p. ex., Smulian, et coll., 2001) et les mesures prises en vue de procurer la formation nécessaire aux personnes préposées à la saisie des données ont donné des résultats positifs (Lorenzoni, et coll., 1999).

3. CONCLUSIONS

Les services de santé représentent une composante importante des dépenses des administrations publiques, des entreprises et des particuliers. La répartition des ressources continue de prendre de l'ampleur et, dans la plupart des provinces et territoires, les services de santé sont le poste le plus important des dépenses des administrations publiques. Afin de procéder aux réformes nécessaires, il est de plus en plus essentiel de disposer des données nécessaires pour éclairer les prises de décision et l'élaboration des politiques. Il est généralement reconnu que les systèmes de santé en place dans les pays comme le Canada doivent être organisés plus efficacement afin de pouvoir répondre aux besoins croissants tout en réduisant les dépenses.

À l'heure actuelle, quelques systèmes d'information seulement emploient des « professionnels des données » pour recueillir, transcrire, gérer et analyser les données sur la santé. La plupart des systèmes existants n'ont jamais été conçus pour appuyer le genre d'analyse que les planificateurs et les décideurs souhaitent réaliser. En outre, fort peu d'attention a été accordée à l'amélioration des processus et les examens de la qualité ont tendance à se concentrer uniquement sur des vérifications ponctuelles de saisie des données. La demande croissante de responsabilisation juridique et financière contribue à faire reconnaître que les organismes de santé doivent mettre en place des procédures en vue d'assurer le traitement correct et uniforme des données (Lichtenstein, et coll., 1999).

Un bon analyste doit acquérir les capacités d'observation d'un détective. Comment les données sont-elles arrivées ici? Qu'a-t-il pu se passer en chemin? Quels motifs auraient pu contribuer à l'introduction d'erreurs systématiques? Avant de pouvoir se fier aux résultats de l'analyse, il est impératif d'estimer la probabilité que les renseignements englobent tous les événements qui devraient l'être, que les données soient enregistrées correctement, que leur saisie ait été faite selon des protocoles cohérents, que les fichiers soient bien construits et complets et que les calculs soient valables et bien décrits. Des erreurs et divers degrés de biais seront indubitablement découverts. Les données doivent donc être analysées et interprétées

en tenant compte des caractéristiques des erreurs. C'est de cette façon uniquement que le détective qui examine les données peut mettre le doigt sur les coupables et amorcer les changements qui amélioreront la qualité des données.

*« Un détective doit tout connaître » (*Sherlock Holmes, *La Vallée de la peur)*

BIBLIOGRAPHIE

Baltimore County Public Library (2001) "Quotes From Sherlock Holmes", http://www.bcpl.net/~lmoskowi/HolmesQuotes.

Bienefeld, M., Woodward G.L. et S. Ardal (2000), "Underreporting of Live Births in Ontario", document non publié, Newmarket, Canada: Central East Health Information Partnership.

Dawson, K. (2000), "Data Quality in RDIS: Issues Related to Combining Data Sets", document non publié, Newmarket, Canada: Central East Health Information Partnership.

Fisher, R.A. (1949), *The Design of Experiments*, Edinburgh: Oliver and Boyd.

Kenney, N. (1999), "Identifying Problems with Data Collection at a Local Level: Survey of NHS Maternity Units in England", British Medical Journal, 319, pp. 619-622.

Lichtenstein, D., Materson, B., et D. Spicer (1999), "Reducing the Risk of Malpractice Claims", Hospital Practice, http://www.hosppract.com/issues/1999/07/licht.htm?3b0bce690.

Lorenzoni, L. Da Cas, R. et Aparo, U.L. (1999), "The Quality of Abstracting Medical Infomraiton From the Medical Record: The Importance of Training Programmes", International Journal of Quality in Helath Care, 11, pp. 209-213.

Smulian, J. C., et Coll., (2001) "New Jersey's Electronic Birth Certificate Program: Variations in Data Sources", American Journal of Public Health, 91, pp. 814-816.

Wolff, N. et Helminiak, T.W. (1996), "Nonsampling Measurement Error in Administrative Data: Implications for Economic Evaluations", Health Economics, 5, pp. 501-512.

Woodward G.L. et Ardal, S. (2000), "Data Quality report: Effect of Residence Code Errors on Fertility Rates", unpublished report, Newmarket, Canada: Central East Health Information Partnership.