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

Examen systématique : évaluation des possibilités de couplage dans les sources de données actuelles

Erin Tanenbaum, Michael Sinclair, Jennifer Hasche, Christina Park[1]

Résumé

La National Children Study, dans sa phase pilote, est une étude de cohorte épidémiologique à grande échelle des enfants et de leurs parents. Des mesures ont été recueillies, de la période précédant la grossesse jusqu'à ce que l'enfant atteigne l'âge adulte. L'utilisation des données existantes visait à compléter la collecte directe de données auprès des répondants. Notre document énonce la stratégie pour le catalogage et l'évaluation des sources de données existantes, en vue de leur utilisation longitudinale à grande échelle. Dans le cadre de notre examen, nous avons sélectionné cinq facteurs d'évaluation pour guider les chercheurs à l'égard des sources de données disponibles : 1) pertinence, 2) actualité, 3) spatialité, 4) accessibilité et 5) exactitude.

Mots-clés : données existantes ; couplage ; données administratives ; évaluation des données.

1. Introduction

1.1 Contexte

La National Children's Study a été conçue pour étudier les influences de l'environnement sur la santé et le développement des enfants. L'étude principale prévue devait étudier 100 000 enfants et leurs parents, de la période précédant la naissance jusqu'à l'âge de 21 ans (NIH, 2011). Le recrutement a pris fin en juillet 2013, et les méthodes mises à l'essai ont profité à d'autres enquêtes longitudinales. Afin de réduire le fardeau de réponse, dans le cadre de la NCS, on prévoyait compléter la collecte de données primaires par une liste de sources de données pouvant être couplées aux données de la NCS par les chercheurs. Il n'est pas inhabituel d'utiliser les sources de données existantes, la science du couplage des fichiers et des enquêtes transversales et longitudinales continuant de progresser. Dans le présent document, nous examinons la nécessité d'outils d'évaluation des sources de données. De tels outils sont nécessaires pour tenir compte des couplages ponctuels, ainsi que pour relier les enquêtes longitudinales à d'autres sources de données longitudinales.

Les sources de données supplémentaires de la NCS portaient sur une vaste gamme de sujets, y compris l'environnement, la santé, l'éducation, la criminalité et les facteurs socioéconomiques. Même si des centaines de sources de données peuvent être couplées aux données de la NCS, la plupart des efforts de recherche analytique pour l'avenir devrait permettre le couplage de seulement quelques fichiers ou sources supplémentaires. La détermination des fichiers de données appropriés peut sembler colossale et, ainsi, dans le présent document, nous résumons les conclusions d'une analyse des ouvrages publiés sur les méthodes d'évaluation des données. Nous commençons par quelques données contextuelles sur l'utilisation des données existantes, puis nous passons à la description d'une structure hiérarchique pour la définition des données existantes, y compris des identificateurs primaires et des critères d'évaluation. En dernier lieu, nous discutons des défis liés aux évaluations longitudinales.

[1] Erin Tanenbaum, NORC, Université de Chicago, NICHD HHSN275201000123U, 4350 East-West Highway, 8e étage, Bethesda, MD, États-Unis, 20814 (Tanenbaum-Erin@norc.org); Michael Sinclair, Mathematica Policy Research, NICHD HHSN275201000123U, 1100 1st Street NE, 10e étage, Washington, DC, États-Unis, 20002-4221 (MSinclair@mathematica-mpr.com); Jennifer Hasche, NORC, Université de Chicago, NICHD HHSN275201000123U, 55 East Monroe Street, 20e étage, Chicago, IL, États-Unis, 60603 (Hasche-Jennifer@norc.org); Christina Park, National Institute of Child Health and Human Development (NICHD), 6100 Executive Blvd., MSC 7510, Bethesda MD, États-Unis, 20892-7510 (parkchris@mail.nih.gov).

1.2 Types de données existantes

Dans le cadre de la NCS, les sources de données existantes comprennent des données concernant des personnes, des groupes de personnes ou des quartiers, dans les cas où les données n'ont pas été recueillies à l'origine par l'entremise de la NCS. Ainsi, les données existantes comprennent diverses sources de données, y compris des registres, des données administratives, des « mégadonnées » et d'autres sources. Même si les chercheurs peuvent habituellement percevoir les organismes gouvernementaux et les programmes statistiques (Iezzoni, 1997) comme des producteurs principaux de telles données, un nombre croissant d'organismes non statistiques sont aussi à l'origine de données existantes. Compte tenu de cela, le terme données existantes englobe maintenant une gamme beaucoup plus vaste de données disponibles, y compris en marketing, en contrôle des stocks, en évaluation de l'exposition et des dommages, en photographie aérienne, en bases de données de navigation, en données administratives, etc. Il serait prématuré de limiter le type de source de données pour une étude longitudinale et multidimensionnelle comme la NCS.

1.3 Justification de l'utilisation et de l'évaluation des données existantes dans le cadre de la NCS

Lorsque l'on utilise des données existantes, les avantages dépassent souvent les inconvénients. Toutefois, il est quand même important de peser soigneusement les avantages et les inconvénients de l'utilisation de données supplémentaires. Dans le cadre de la NCS, on prévoyait dépendre dans une large mesure des fichiers de données existantes pour les données au niveau de la région et du répondant.

Du point de vue de la collecte, les sources supplémentaires sont souvent moins coûteuses à acquérir que les données primaires (Iezzoni, 1997), sont facilement disponibles, peuvent englober de grandes populations et peuvent réduire le fardeau pour les répondants. En outre, certaines sources externes peuvent constituer la meilleure source d'information sur un sujet, ce qui rend irrationnelle la collecte dans le cadre d'une étude primaire (p. ex., un recensement pour des estimations de la population). Les données existantes peuvent aussi être utilisées par les statisticiens pour améliorer la couverture des bases d'échantillonnage, corriger les biais dans l'imputation, permettre un contrôle direct des questions d'enquête et vérifier l'exactitude ou l'uniformité des réponses aux enquêtes (Chappell, 2005, Bradburn, 1993, Czajka, 2003).

Dans le cas des chercheurs, les données existantes peuvent combler des lacunes lorsque les études ne permettent pas de prédire tous les besoins de données ou de recueillir toutes les données nécessaires. En outre, l'utilisation de données administratives sur la santé continue de prendre de l'expansion comme façon de révéler et d'améliorer peut-être la santé des personnes (Daver, 2013, Ng, 2010, Zhan, 2003). Par exemple, un patient de l'assurance-maladie saura sûrement qu'il a eu un « pontage coronarien », mais ne pourra probablement pas rendre compte de l'intervention effectuée et des coûts connexes, les demandes de remboursement pouvant combler cette lacune, ce qu'une enquête à elle seule ne peut pas faire. Ainsi, dans le riche environnement de données d'aujourd'hui, la question de déterminer si un chercheur devrait utiliser les données existantes a changé, et on se demande plutôt quelles données existantes devraient être utilisées et quelles sont les limites de cette source.

1.4 Avantages d'une évaluation

Même si l'utilité des données existantes est claire, des défis ont été notés. Par exemple, certains peuvent coupler de façon inappropriée les données existantes, ou laisser de côté les préoccupations concernant la qualité des données (Iezzoni, 1997, Jabine, 1985). En outre, leur utilisation peut entraîner une mauvaise utilisation par inadvertance. Par exemple, lorsque les médecins utilisent les dossiers de santé électroniques (EHR), ils peuvent assurer le codage en utilisant des exigences, des procédures, des degrés de rigueur ou d'exactitude différents (Iezzoni, 1997), ce qui fait en sorte qu'il est difficile de dire si les différences dans les dossiers administratifs sont le reflet de différences véritables. Par exemple, les certificats de naissance sont souvent limités quant au nombre de catégories pour le type d'assurance et de changements de catégorie selon l'état (Martin, 2013). En outre, les données existantes manquent souvent de documentation. Idéalement, une documentation rigoureuse et uniforme au fil des ans devrait être mise à la disposition des chercheurs, mais cela est rarement le cas, ce qui rend les évaluations de données extrêmement difficiles à effectuer (Reidy, 1998).

Par ailleurs, les préoccupations en matière de protection des renseignements personnels continuent d'augmenter. Il semble que plus (et non pas moins) de données font l'objet de restrictions quant à l'accès qui nuisent aux avantages possibles (Lane, 2010). Ainsi, un mécanisme qui a été annoncé comme une source de données plus rapide et moins

coûteuse (Banque mondiale) comporte des coûts, y compris la possibilité de tirer des inférences incorrectes (Davern, 2013), ou de présumer que l'accès aux données est disponible, alors qu'il ne l'est pas dans les faits. Toutefois, il peut être extrêmement avantageux de coupler des sources à accès restreint. Par exemple, le couplage au moyen de données identifiables sur des personnes, comme le nom, la date de naissance, l'adresse et le numéro de sécurité sociale, peut réduire le fardeau pour les répondants et augmenter les connaissances des chercheurs, particulièrement lors du couplage de données inconnues pour le répondant (p. ex., codes de diagnostic). La possibilité d'évaluer les données existantes est par conséquent primordiale, compte tenu des défis et des avantages possibles liés à l'utilisation des sources de données existantes. Toutefois, on doit répondre à la question suivante : l'amélioration analytique compense-t-elle les coûts du couplage des données existantes ? C'est pourquoi, une bibliothèque de données existantes et une méthode d'évaluation ont été élaborées pour la NCS.

2. Évaluation des données existantes

2.1 Une approche multifactorielle

Nous avons effectué un examen des ouvrages publiés concernant les méthodes d'évaluation des données, en commençant par des normes bien établies à partir des ouvrages publiés dans les domaines médical, statistique et des enquêtes sociales. Après avoir passé en revue plus de 80 publications, nous avons recensé les concepts d'évaluation selon la source et compilé les concepts en catégories. Après un examen initial, nous nous sommes rendu compte que l'examen des concepts individuellement ne tient pas compte des nuances des ouvrages publiés.

Certains auteurs ont choisi une évaluation structurée liée au type d'information concernant les données. Par exemple, plusieurs auteurs ont déterminé des aspects de qualité liés : 1) aux sources de données (conservation et fourniture de la source de données), 2) aux métadonnées (clarté de la définition des données, etc.), et 3) aux données proprement dites (données probantes comprises dans la source de données) (ABS, 2006). Karr, Sanil et Banks (2006) ont poussé le concept plus loin en classant les trois selon un niveau croissant de détail nécessaire, selon l'hypothèse qu'un rapport sur la qualité des données serait plus détaillé que ceux sur les métadonnées et la source. Les auteurs visaient par ce tri à permettre à un utilisateur de cesser de chercher de l'information si un aspect était considéré comme non pertinent, ou de laisser de côté un système de données, s'il échouait à un niveau supérieur.

À partir de notre examen, nous recommandons une approche multifactorielle comportant deux facteurs : une hiérarchie ponctuelle et une hiérarchie conceptuelle. Une telle hiérarchie peut réduire le niveau d'effort si les résultats initiaux concernant la source de données révèlent qu'elle ne convient pas à l'utilisation prévue. Par exemple, les chercheurs ne voudraient pas coupler le taux de vaccination au niveau national tiré d'une enquête si 97 % des répondants n'avaient pas répondu aux questions sur la vaccination. Dans un tel exemple, le chercheur pourrait cesser d'évaluer la source et chercher d'autres fichiers pour obtenir des données similaires. Notre système s'appelle Protocole d'évaluation approprié, les besoins particuliers du chercheur étant au centre de l'évaluation.

2.2 Protocole d'évaluation approprié

Afin d'ancrer notre bibliothèque de données existantes, nous avons commencé par créer un cadre. Une bibliothèque de données existantes a été créée pour aider les chercheurs à déterminer les sources de données possibles pouvant être utilisées comme filtres primaires. Cette bibliothèque pourrait être perçue comme le premier outil dans notre protocole proposé d'évaluation des utilisations appropriées. Toutefois, l'examen de centaines de sources de données est à la fois peu pratique et peu raisonnable pour la plupart des efforts de recherche. Ainsi, des filtres ont été créés pour permettre aux chercheurs de filtrer rapidement les sources disponibles. Les filtres de la bibliothèque de données existantes de la NCS comprenaient le nom de la base de données, le fournisseur des données, le sujet, les sujets auxiliaires et les éléments de données clés. D'autres données ont aussi été recueillies sur chaque source, y compris le type de fournisseur, les données de contact (site Web, etc.) et une liste de bases de données similaires permettant aux chercheurs de passer en revue rapidement les sources disponibles, avant de passer à la couche suivante de notre protocole d'évaluation proposé : les critères d'évaluation.

À partir des ouvrages publiés, nous avons passé en revue les critères d'évaluation avec l'intention de les intégrer dans notre bibliothèque de données existantes. Toutefois, nous avons déterminé que nombre des idées étaient impossibles à mettre en œuvre. Par exemple, l'« exactitude » a été souvent mentionnée, mais un utilisateur des données ne peut pas mesurer l'exactitude sans d'autres instructions. Ainsi, nous avons élaboré cinq facteurs de base qui ont trait aux 27 concepts (les questions qui sous-tendent les facteurs), et 50 éléments pouvant faire l'objet d'un suivi (la mesure

derrière le concept). Les éléments peuvent quant à eux être résumés pour créer un aperçu rapide, afin d'évaluer une source pour un facteur ou un concept donné.
Il est important de noter qu'une vaste évaluation peut ne pas être obligatoire avant l'utilisation d'une source. Parfois, l'expérience antérieure fournit suffisamment de renseignements pour sélectionner un fichier de données existantes. Néanmoins, un système d'évaluation fournit une façon structurée d'étudier rapidement la qualité dans son intégralité ou par facteur pour une application particulière ou une source possible de nouvelles données.
Nos facteurs recommandés d'évaluation de la bibliothèque sont similaires à ceux figurant dans les ouvrages publiés (Kasprzyk, 2001), sauf pour une exception. Nous avons divisé l'actualité en attributs temporels et géographiques. La distinction est importante dans le cadre de la NCS, les sources de données environnementales jouant un rôle clé et la période de vie des produits chimiques variant considérablement (Lioy, 2009) et les distances recommandées entre les sources de pollution et les résidences variant aussi. En séparant le temps et l'espace, une source peut répondre à nos besoins géographiques, tout en étant périmée ou recueillie selon un intervalle irrégulier.

Nos cinq facteurs contribuent à répondre aux questions suivantes :

1. Accessibilité : quel niveau de difficulté présentent l'acquisition et l'utilisation des données ?
2. Exactitude : les mesures sont-elles documentées du point de vue de l'exactitude, de la précision et/ou de l'uniformité entre les régions géographiques ?
3. Pertinence : les données servent-elles les objectifs analytiques ? La source fournit-elle des facteurs confusionnels ou des résultats qui sont difficiles à recueillir dans le cadre du système de collecte de la NCS?
4. Spatial : à quel niveau de spécificité géographique sont-elles disponibles ?
5. Actualité : quand les données ont-elles été recueillies et avec quelle rapidité sont-elles distribuées ?

Après avoir déterminé 50 éléments pouvant faire l'objet d'un suivi, nous les avons répartis en volets séquentiels : le volet 1 peut être utilisé par toutes les études, comme outil de filtrage. Le volet 2 comprend les éléments uniquement disponibles avec accès aux données ou en rapport avec une étude. Une fois les éléments du volet 1 recueillis, un chercheur peut coter les éléments selon ses propres besoins et décider des sources de données à explorer davantage. Les tableaux 2.2-1 et 2.2-2 comprennent une liste complète des facteurs, composantes et types de réponse proposés.
Même si ces critères d'évaluation sont utiles, il ne suffit pas d'évaluer simplement une base de données. La majeure partie des ouvrages publiés ont souligné que l'évaluation des données dépend de l'utilisation prévue. Par exemple, une étude des populations urbaines peut laisser de côté le fait que la couverture des régions rurales est limitée dans une source de données. Comme les besoins de données varieront selon la question de recherche, un mécanisme de pointage final ne fait pas partie de notre liste d'éléments. Même si certaines sources ont proposé une fiche de classement, nous croyons que les renseignements concernant chaque élément devraient être recueillis, et que l'importance devrait par la suite être déterminée par le chercheur, et non pas un bibliothécaire de données existantes. Par ailleurs, ce ne sont pas tous les éléments qui peuvent être recueillis et, ainsi, des sources de données de filtrage peuvent être nécessaires avec les données limitées. Dans ces cas, nous croyons que les chercheurs pèseront les avantages et les inconvénients des données manquantes.

Tableau 2.2-1
Critères du volet 1 pour l'ensemble des études

Facteurs	Concepts	Éléments	Réponses
Accessibilité	Facilité d'accès aux données ?	Données disponibles dans le public?	Oui/non.
		Disponibles par l'entremise d'un site Web?	Oui/non.
		Accès aux instructions de données	URL ou brève description.
		Restrictions concernant l'utilisation ou la diffusion?	Oui/non
		Délai requis pour recevoir un fichier après la date de la demande.	Nombre de jours ou de mois.

Facteurs	Concepts	Éléments	Réponses
	Documentation suffisante ?	Fichiers disponibles, y compris : 1) questionnaire, 2) méthodologie, 3) rapport sur la qualité, 4) tableau de concordance entre les versions de produit de données, et 5) fichier des codes?	Oui/non pour chacun; URL si disponible.
		Documentation appropriée (y compris langue, symboles, unités, etc.)?	Liste de vérification pour chaque type de documentation.
	Bon rapport coût-efficacité?	Coût d'achat des données	En dollars.
	Logiciels standard disponibles?	Quels sont les logiciels utilisés? (SAS, ASCii, Microsoft Access, etc.).	Oui/non pour chacun.
Exactitude	Couplage possible?	Présence d'un identificateur unique pour le couplage.	Oui/non. Plus instructions.
	Non-réponse unitaire?	Pourcentage d'unités dans lesquelles toutes les données sont manquantes.	Taux.
	Précision?	Erreur-type, erreur d'échantillonnage disponible?	Oui/non ou s/o (recensement).
	Contrôles appropriés?	Imputations, pondération, contrôles de vérification logique, corrections publiées après la diffusion, etc.	Oui/non pour chacun.
Pertinence	Utilisation des données appropriée?	Méthode de collecte des données appropriée?	Liste des méthodes de données et brève explication concernant leur pertinence.
	Réputation	Les données sont-elles considérées comme vraies et crédibles?	Score, nombre de citations utilisant l'ensemble de données. Points négatifs si la citation trouve des défauts en ce qui a trait à la source.
	Couplage approprié?	Variable de couplage prévue appropriée?	Cote (1 à 10) par rapport aux besoins de l'étude.
		Comparabilité des définitions de l'unité de couplage?	Explication écrite des différences.
		Couplage validé? Couplé précédemment (autres études)?	Oui/non.
	Représentation uniforme?	Définitions de données et méthodologies stables au fil du temps?	Score de stabilité (1 à 10) et aperçu des changements.
Aspect spatial	Niveau géographique disponible?	Listé comme a) accès public, b) accès restreint, c) non disponible	Réponses pour chaque région géographique, y compris nationale, état, régionale, comté, secteur, îlot, ménage, personne, etc.
		Les données sont-elles complètes?	Oui/non pour chaque niveau.
	Données GPS	Coordonnées disponibles ou faciles à calculer?	Oui/non et niveau.
Actualité	Compatibilité temporelle	Fréquence de la collecte des données	Fréquence (mensuelle, etc.).
		Période de la collecte des données?	Mois/année de début jusqu'à la date prévue.
		Délai pour la diffusion à partir de la collecte des données.	Période.

Facteurs	Concepts	Éléments	Réponses
		Documentation temporelle suffisante?	Oui/non.

Tableau 2.2-2
Volet 2 Critères dans une étude

Facteurs	Concepts	Éléments	Réponses
Accessibilité	Codes de programme d'analyse fournis?	Programmes supplémentaires pour importer et/ou analyser les données.	Oui/non selon le type de données.
	Coût de traitement des données?	Coûts estimés pour formater les données.	En dollars.
Exactitude	Intégralité	Portée des données manquantes.	% de données manquantes selon l'élément de données.
		Champs imputés?	% de données imputées, et documentation présente ou non pour les méthodes.
	Précision uniforme?	Examiner les changements dans les taux de réponse, l'attrition, etc., au fil du temps.	Évaluation (1 à 10).
	Étendue des données?	Couverture de la population cible?	% de couverture ou évaluation (1 à 10).
	Erreur de mesure?	Énoncés concernant les biais.	Données ouvertes; évaluation (1 à 10).
Pertinence	Uniformité de la structure logique	Tests des rapports logiques.	Séquence présente ou non, modèles d'enchaînement, vérification de la cohérence, etc.
	Utilisation des données appropriée?	Mots-clés liés au contenu nécessaire?	Liste de textes à partir de répertoires de codes ou de sites Web de sources.
		Utilité des éléments de données?	Liste des éléments des répertoires de codes. Évaluation (1 à 10) selon l'élément.
		Structure logique appropriée?	Oui/non si les tests sont appropriés.
		Source des données appropriée pour l'étude?	Catégorique (gouvernement, secteur privé, autres); évaluation (1 à 10).
	Population cible compatible?	Passer en revue la population incluse dans les données par rapport à la population de la NCS.	Score de 1 à 10. Comprend les concepts, comme les logements collectifs, la définition d'un ménage, etc.
	Utilité, valeur ajoutée, utilisation par d'autres	Selon le but visé. Les données servent à compléter/ remplacer la collecte directe?	Propre à la recherche.
		Données utilisées par d'autres?	Nombre de citations.
Actualité	Compatibilité temporelle	Fréquence correspondant aux besoins de l'étude?	Évaluation de 1 à 10 pour l'utilisation prévue.
		Période des données correspondant aux besoins de l'étude?	Évaluation de 1 à 10 pour l'utilisation prévue.

Les études longitudinales comme la NCS comportent des défis uniques du point de vue du couplage des données, des sources de données, des méthodes de collecte et des modifications du niveau de documentation au fil du temps. Malheureusement, il est souvent difficile de déterminer et de localiser la documentation des données de nombreuses années plus tard; toutefois, les analyses sont souvent financées ou conceptualisées des années après la collecte des données. Par exemple, imaginons qu'une analyse soit prévue deux ans après la collecte des données. À ce moment-là, les sources de données existantes sont évaluées et couplées selon la recherche. Toutefois, l'équipe qui a créé les données existantes peut changer, la compagnie qui fournit les données peut fermer ses portes ou fusionner, et la documentation peut être manquante. Idéalement, de la documentation complète devrait être mise à la disposition du chercheur lorsque les données sont utilisées pour l'analyse, mais cela est rarement le cas, ce qui rend l'évaluation des données existantes extrêmement difficile à effectuer après coup (Reidy, George, et Lee, 1998).

3. Conclusion

Nous proposons que les chercheurs utilisent un protocole d'évaluation approprié pour identifier rapidement les données existantes qui conviennent pour répondre à leurs questions de recherche individuelles. Le présent document comprend un sommaire des critères d'évaluation que nombre de personnes ont utilisé pour étudier l'utilisation appropriée de sources de données existantes particulières, en vue de compléter les données de programme, selon l'hypothèse que ces données amélioreront l'utilité analytique et la qualité des résultats de l'étude et peuvent réduire les coûts de collecte des données et le fardeau. Ces critères fournissent une façon de prendre des décisions à l'intérieur d'un ensemble concurrent de sources de données existantes, et nous fournissons une infrastructure suggérée pour la collecte et l'entreposage de ces données à divers niveaux d'examen. Nous souhaitons que ces méthodes aident d'autres chercheurs qui font face à des défis similaires à concevoir leur propre programme de couplage de données.
Étant donné que la qualité de l'analyse correspond à celle des données sur lesquelles elle est fondée et que la qualité des données dépend du processus qui sert à leur collecte, en tant qu'auteurs, nous souhaitons un environnement dans lequel la transparence des données répond aux besoins de tous les chercheurs. En attendant, nous continuerons de chercher des critères d'évaluation des données.

Bibliographie

Australian Bureau of Statistics (ABS). (2006). Information Paper: Evaluation of Administrative Data Sources for Use in Quarterly Estimation of Interstate Migration, 2006 to 2011. (Cat. no. 3127.0.55.001). Canberra, Australia.

Bradburn, N.M. (1993). A Census that Mirrors America [electronic resource]: Interim Report / Panel to Evaluate Alternative Census Methods, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press.

Chappell, G., Obenski, S., and Farber, J. (2005). Research to Improve Census Imputation Methods: Item Results and Conclusions. Presentation at the Joint Statistical Meetings of the American Statistical Association, Survey Research Methods Section. Minneapolis, MN, August 10.

Czajka, J.L., Jacobson, J.E., & Cody, S. (2003). Survey estimates of wealth: a comparative analysis and a review of the Survey of Income and Program Participation. Social Security Bulletin, 65(1), 63.

Davern, M., Roemer, M., and Thomas W. (2013). Investing in a Data Quality Research Program for Administrative Data Linked to Survey Data for Policy Research. Unpublished book.

Iezzoni, L. (1997). Assessing quality using administrative data. Annals of Internal Medicine, 127(8 Pt 2), 666-674.

Jabine, T.B., & Scheuren, F.J. (1985). Goals for statistical uses of administrative records: the next 10 years (with discussion). Journal of Business and Economic Statistics, 3, 380-404.

Karr, A.F., Sanil, A.P., & Banks, D.L. (2006). Data quality: a statistical perspective. Statistical Methodology, 3(2), 137. doi:10.1016/j.stamet.2005.08.005

Kasprzyk, Daniel (2001). Talk at the National Statistics Office (NSO) at the Federal Committee on Statistical Methodology (FCSM) conference. http://www.fcsm.gov/01papers/Kasprzyk.pdf Accessed July 1, 2013.

Lane, J., & Schur, C. (2010). Balancing access to health data and privacy: a review of the issues and approaches for the future. Health Services Research, 45(5 Pt 2), 1456-1467. doi:10.1111/j.1475-6773.2010.01141.x

Lioy, P., Isukapalli, S., Trasande, L., Thorpe, L., Dellarco, M., Weisel, C., & ... Landrigan, P. (2009). Using national and local extant data to characterize environmental exposures in the National Children's Study: Queens County, New York. Environmental Health Perspectives, 117(10), 1494-1504. doi:10.1289/ehp.0900623

Martin, J. A., Wilson, E. C., Osterman, M. J., Saadi, E. W., Sutton, S. R., & Hamilton, B. E. (2013). Assessing the quality of medical and health data from the 2003 birth certificate revision: results from two states. National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System, 62(2), 1-19.

National Institutes of Health (NIH), The National Children's Study: It's All About Our Children. NIH MedlinePlus: the magazine [Internet]. 2011 Summer;6(2):4-5. Available from: https://www.nlm.nih.gov/medlineplus/magazine/issues/summer11/articles/summer11pg4-5.html

Ng, C. (2010). Population and Administrative Datasets for Research and Evaluation. Presentation for Fraser Health.

Reidy, M., George, R., & Lee, B.J. (1998). Developing an integrated administrative database. Exploring Research Methods in Social Policy Research. Asldershot, UK: Ashgate Publishing Company.

The World Bank, What Happens When Big Data Meets Official Statistics? - Live Webcast. Speakers include Robert Groves, Provost, Georgetown University; formerly Director, U.S. Census Bureau. Accessed from: http://live.worldbank.org/what-happens-when-big-data-meets-official-statistics-live-webcast

Zhan, C., & Miller, M. (2003). Administrative data-based patient safety research: a critical review. Quality & Safety in Health Care, 12(Suppl 2), ii58-ii63.