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

Applications pratiques du calcul sécurisé pour le contrôle de la divulgation

Luk Arbuckle, Khaled El Emam[1]

Résumé

La diffusion de microdonnées exige habituellement des méthodes de réduction et de modification des données, et le degré d'application de ces méthodes dépend des méthodes de contrôle qui seront nécessaires pour accéder aux données et les utiliser. Le calcul sécurisé est une approche qui, dans certaines circonstances, convient davantage pour accéder aux données à des fins statistiques; il permet le calcul de fonctions analytiques à l'égard de données chiffrées sans qu'il soit nécessaire de déchiffrer les données sources sous-jacentes pour procéder à une analyse statistique. Cette approche permet aussi à plusieurs emplacements de fournir des données, tout en garantissant une protection rigoureuse de la vie privée. De cette façon, les données peuvent être regroupées, et les fournisseurs de données peuvent calculer des fonctions analytiques, sans qu'aucune des parties ne connaisse les entrées des autres. À l'aide de certains résultats théoriques et d'exemples réels issus du domaine des soins de santé, nous expliquerons comment le calcul sécurisé peut être appliqué dans des contextes pratiques.

Mots clés : Contrôle de la divulgation, calcul sécurisé, analyse à distance.

1. Introduction

Dans le cadre d'une approche fondée sur les risques en matière de contrôle de divulgation, le calcul sécurisé peut être envisagé en tant que données pseudonymes « protégées ». Le chiffrement assure la « protection », du fait qu'aucun humain n'est en contact direct avec les données : les utilisateurs voient plutôt uniquement les résultats statistiques pour des variables clés. Même si les données sources sont protégées par chiffrement lors du calcul sécurisé, certaines préoccupations concernant les systèmes d'analyse à distance en général demeurent et méritent qu'on s'y attarde. Le calcul sécurisé convient bien aux scénarios caractérisés par une analyse et une collecte continues et systématiques de données, parce que les calculs peuvent, dans de tels cas, être définis et optimisés, puis appliqués de façon continue.

1.1 Contrôle de la divulgation en fonction du risque

Ce n'est pas d'hier que les méthodes de contrôle de la divulgation statistique tiennent compte du rapport (distribution marginale) aux variables clés ou aux quasi-identificateurs pour protéger contre une divulgation de l'identité (Duncan et coll. 2011). Les règlements et les documents d'orientation préconisent l'utilisation d'un cadre de désidentification qui est fondé sur le risque, c.-à-d. qui incorpore le contexte de diffusion des données au cadre d'évaluation du risque, afin d'obtenir de solides garanties que le risque est « raisonnable ».

Nous supposerons que le masquage des identificateurs uniques ou directs est bien compris et axerons plutôt notre propos sur les variables clés. Considérons la probabilité de réidentification dans l'éventualité où un attaquant ferait la tentative suivante $\Pr(\text{reid} \mid \text{attempt})$ (Marsh et coll. 1991). De là, nous pouvons formuler le problème comme correspondant à la probabilité d'une réidentification et d'une attaque à l'aide de l'équation suivante :

$$\Pr(\text{réid}, \text{tentative}) = \Pr(\text{réid} \mid \text{tentative}) \times \Pr(\text{tentative}).$$

La probabilité qu'un attaquant tente une réidentification est déterminée d'après le contexte de diffusion des données, à l'aide d'une évaluation subjective du risque (Morgan et coll. 1992; Vose 2008) fondé sur l'opinion d'experts et sur les précédents (p. ex. (Centers for Disease Control and Prevention 2004; Statistique Canada 2007; Subcommittee on Disclosure Limitation Methodology 2005). Les facteurs qui influent sur une attaque comprennent les pratiques du

[1]Luk Arbuckle, Institut de recherche du Centre hospitalier pour enfants de l'est de l'Ontario, 401, Smyth Road, Ottawa (Canada), K1H 8L1; Khaled El Emam, Institut de recherche du Centre hospitalier pour enfants de l'est de l'Ontario, 401, Smyth Road, Ottawa (Canada), K1H 8L1, et Faculté de médecine de l'Université d'Ottawa, 451, Smyth Road, Ottawa (Canada), K1H 8M5.

demandeur de données en matière de sécurité et de respect de la vie privée, ainsi que les obligations contractuelles. De plus, un seuil de risque défendable peut être établi à la lumière des précédents en évaluant la possibilité d'une atteinte à la vie privée.

Tous les facteurs susmentionnés peuvent être utilisés pour établir un cadre d'évaluation du risque reproductible (El Emam 2013). En fait, nombreuses sont les normes et les lignes directrices qui ont intégré un tel cadre de travail fondé sur le risque à leurs recommandations concernant le partage des données sur la santé (Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine. 2015; Health Information Trust Alliance 2015; Information Commissioner's Office 2012; Office for Civil Rights 2012; PhUSE De-Identification Working Group 2015; le comité d'experts sur l'accès en temps opportun aux données sur la santé et sur les conditions sociales pour la recherche sur la santé et l'innovation du système de santé 2015).

1.2 Calcul sécurisé

Supposons que de multiples parties souhaitent regrouper des données et calculer une fonction sans qu'aucune d'entre elles ne connaisse les entrées des autres. L'idée fondamentale sur laquelle repose le calcul sécurisé est de pouvoir calculer une fonction sur des données chiffrées, sans qu'il soit nécessaire de chiffrer les données pour obtenir la sortie désirée. Les primitives cryptographiques, ou éléments primaires, utilisés pour créer des protocoles de calcul sécurisé peuvent provenir du chiffrement homomorphe, de circuits brouillés, du partage de secret ou autres, qui présentent tous des avantages et des inconvénients qui leur sont propres.

Le chiffrement homomorphe est devenu applicable grâce à l'avènement du système de chiffrement de Paillier, qui a rendu le temps de calcul raisonnable; il comporte cependant un ensemble limité d'opérations (Paillier 1999). Les exemples pratiques que nous présentons dans les sections qui suivent sont axés sur cette technologie. Les circuits brouillés de Yao (Yao 1986) ont été considérés comme difficilement applicables en raison du temps de calcul nécessaire et de la capacité de mémoire requise; l'avènement de nouvelles méthodes pourrait cependant changer la donne (Gueron et coll. 2015; Songhori et coll. 2015). Un schéma de partage de secret tel que celui de Shamir (Shamir 1979) serait efficace du point de vue des calculs à l'intérieur d'un système homomorphe (Benaloh 1986). Certains se sont dits préoccupés par le fait que les parties doivent révéler leurs parts au moment de la reconstruction du secret (c.-à-d. pour obtenir la sortie finale). Dans de tels cas, d'autres techniques pourraient être utilisées (Desmedt et Frankel 1989), ou les applications pourraient répartir au hasard de nouvelles parts représentant la même valeur de secret sans divulguer le résultat lui-même.

Le domaine du calcul sécurisé progresse rapidement grâce à des méthodes plus efficaces et plus échelonnables (Rohloff et Cousins 2014), ainsi qu'à des composantes informatiques spécialisées qui permettent d'accélérer les calculs (Cousins et coll. 2015). En outre, des logiciels d'analyse statistique polyvalents utilisant des systèmes de chiffrement sont également développés (Dan Bogdanov et coll. 2014).

Le calcul sécurisé rejoint les directives établies par les organismes de réglementation, à titre de moyen permettant à la fois de protéger les renseignements personnels sur la santé et de partager des données chiffrées à des fins d'analyse collaborative. Le calcul sécurisé implique également les plus hauts niveaux de contrôles de sécurité, car aucun traitement de renseignements personnels sur la santé n'est effectué par l'humain. Moyennant des obligations contractuelles et des mesures garantissant l'absence de fuites de renseignements personnels à partir des résultats eux-mêmes (O'Keefe et Chipperfield 2013), le calcul sécurisé peut être envisagé, dans le contexte d'un cadre de travail fondé sur le risque, comme correspondant à des données pseudonymes *protégées* assorties d'un risque très faible de réidentification.

2. Applications pratiques

2.1 Couplage sécurisé

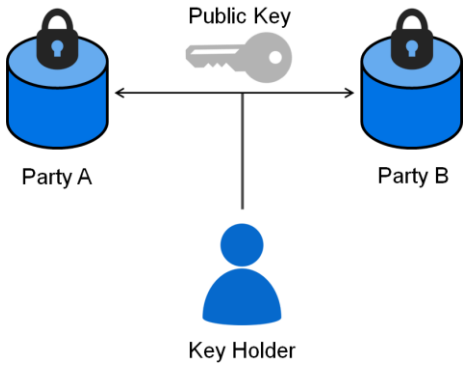
Le couplage peut être utilisé pour procéder à un examen des enregistrements ou à un appariement de bases de données dans le but d'éliminer les enregistrements en double. Or, les meilleurs champs pour effectuer un couplage sont souvent ceux qui ne peuvent pas être divulgués (p. ex., numéro d'assurance sociale, prénom et nom de famille). Le but du

couplage sécurisé est de coupler sans partager de renseignements personnels ou sensibles, et cela ne se limite pas uniquement aux champs utilisés pour le couplage. Le fait de révéler à une autre partie quels sujets de données sont contenus dans une base de données peut constituer une divulgation en soi si la liste des membres de la base de données est une information sensible.

Dans le cadre de notre protocole de couplage (El Emam et Arbuckle 2013 chap. Secure Linking), nous avons recours à des calculs sécurisés impliquant un tiers semi-sécurisé à qui l'on fait confiance pour exécuter le protocole, mais qui serait incapable d'obtenir des renseignements personnels ou sensibles concernant les sujets des données même s'il le voulait. Parce que les données et les calculs effectués sur les données sont protégés par chiffrement, il n'est pas nécessaire que les parties se fassent mutuellement confiance ou qu'elles fassent confiance au tiers semi-sécurisé. Aucune des parties concernées ne peut « piocher » dans les données ou les calculs. Qui plus est, une intrusion à l'un ou l'autre des emplacements n'aurait pas pour effet de révéler l'identité ou les renseignements personnels des sujets des données.

Tel qu'illustré à la Figure 2.1-1, la première étape de notre protocole de couplage sécurisé consiste à ce que des clés privées et publiques soient générées par le détenteur de la clé, puis à distribuer la clé publique aux gardiens des données afin qu'ils puissent l'utiliser pour chiffrer les variables de couplage. Les données chiffrées seront ensuite transmises à un agrégateur central ou, subsidiairement, envoyées à un gardien des données par un autre gardien des données. Une vérification de l'égalité homomorphe des données chiffrées regroupées est ensuite effectuée. Les résultats d'appariement chiffrés sont ensuite envoyés au détenteur de la clé, qui les déchiffre à l'aide de la clé privée.

Figure 2.1-1
Distribution de la clé publique aux gardiens des données



Ce protocole sécurisé est utilisé par l'Institute for Clinical Evaluative Sciences (ICES) pour le couplage de données désidentifiées (appariées selon le numéro d'assurance sociale, le nom et la date de naissance), et a été proposé pour déterminer les taux de dépistage et de test pour la chlamydia en collaboration avec un organisme de santé publique (en appariant les dossiers médicaux électroniques détenus par les médecins de famille avec les résultats des tests de laboratoire). Il a également été proposé aux fins d'une étude d'impact d'une initiative de vaccination contre le virus du papillome humain (VPH), laquelle fournit d'ailleurs davantage de renseignements au sujet du protocole (El Emam et coll. 2012a).

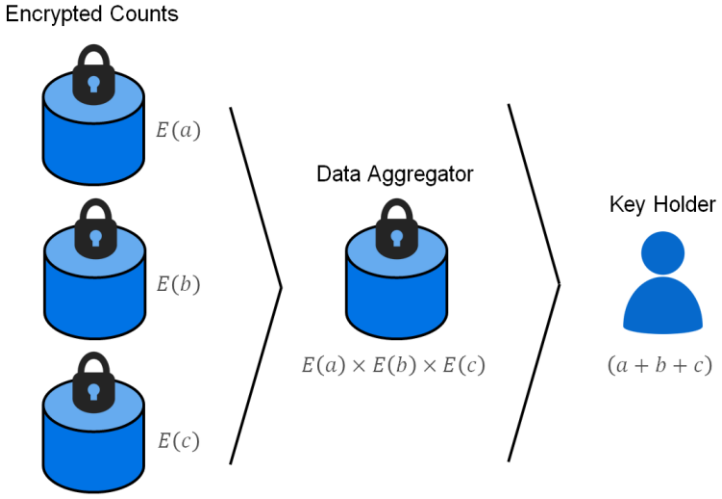
2.2 Prévalence des organismes résistant aux antimicrobiens

Grâce à un système de collecte de données sécurisé, qui procure de solides garanties en matière de confidentialité et de respect de la vie privée, nous avons pu réaliser une étude de la prévalence ponctuelle pour évaluer les taux d'organismes résistant aux antimicrobiens (ORA) dans les établissements de soins de longue durée en Ontario (El Emam et coll. 2014). L'identification des résidents porteurs d'ORA dans les établissements de soins de longue durée peut être source de stigmatisation, mais grâce à des calculs sécurisés, des données sur la colonisation et l'infection ont pu être recueillies sans révéler les taux observés dans les établissements participants. Cette étude a répondu à la nécessité de recueillir des données sur la prévalence des ORA dans les établissements de soins de longue durée à des fins de surveillance et d'intervention en santé publique.

Le cadre de base du système de collecte de données sécurisé est illustré à la Figure 2.2-2, dans laquelle $E(m)$ représente le texte chiffré qui remplace le texte en clair m, et dans laquelle la multiplication du texte chiffré selon le système de chiffrement de Paillier est équivalente à l'addition du texte en clair, c'est-à-dire que $E(a) \times E(b) = E(a + b)$. Les établissements de soins de longue durée ont fourni les dénombrements, qui ont été chiffrés au point de collecte. Ces dénombrements chiffrés ont ensuite été combinés par un agrégateur de données à l'aide de calculs sécurisés, lequel agrégateur a ensuite transmis les résultats intermédiaires au détenteur de la clé pour déchiffrer. Des calculs sécurisés ont été utilisés pour déterminer les taux moyens de colonisation ou d'infection et l'écart-type, selon la région ou la taille de l'établissement, et pour exécuter un test de randomisation à deux échantillons (test t randomisé) à l'égard du biais de non-réponse.

On a demandé à tous les établissements de soins de longue durée de la province de fournir les chiffres de colonisation ou d'infection pour le staphylocoque doré résistant à la méthicilline (SARM), les entérocoques résistants à la vancomycine (ERV), et les bactéries productrices de bêta-lactamase à spectre étendu (BLSE) consignés dans leurs dossiers médicaux électroniques, ainsi que le nombre actuel de leurs résidents. Les données ont été recueillies en ligne pendant la période d'octobre et novembre 2011. Au total, 82 % des établissements de la province ont répondu, ce qui représente un taux de réponse beaucoup plus élevé que lors des tentatives de collecte de données précédentes (effectuées sans le recours au calcul sécurisé). Les constatations microbiologiques et leur répartition se sont avérées conformes aux données des laboratoires provinciaux disponibles concernant les résultats du dépistage des ORA dans les hôpitaux.

Figure 2.2-2
Calcul de la prévalence ponctuelle à l'aide d'un système de chiffrement de Paillier



2.3 Événements iatrogènes médicamenteux rares

La régression logistique est couramment utilisée pour l'analyse des événements iatrogènes médicamenteux. Les données doivent cependant être regroupées pour permettre la détection des événements rares et assurer une hétérogénéité suffisante de la population afin de garantir l'innocuité et l'efficacité d'un médicament chez les sous-populations. Nous avons donc mis au point un protocole de régression logistique distribuée sécurisée n'utilisant qu'un seul centre d'analyse relié à de multiples emplacements fournisseurs de données (un peu comme dans l'exemple précédent) à des fins de surveillance post-commercialisation. Nous avons également élargi le protocole pour utiliser des équations d'estimation généralisées (GEE) afin de tenir compte des données corrélées, d'autres modèles linéaires généralisés (GLM), et des modèles de survie (El Emam et coll. 2012b).

Pour estimer un modèle linéaire généralisé (Agresti 2002), on peut utiliser la méthode de Newton-Raphson et calculer itérativement les estimations des paramètres b à l'aide de l'équation

$$b(t+1) = b(t) - [I(t)]^{-1}u(t),$$

dans laquelle $u(t)$ correspond au vecteur de score et $I(t)$ est la matrice d'information pour l'itération de t. Lorsque de multiples emplacements (partitionnés horizontalement afin qu'ils aient les mêmes covariables et formats de codage) fournissent des données, le vecteur de score et la matrice d'information sont, en fait, calculés individuellement à chaque emplacement, puis combinés ultérieurement. Ce qui correspond, dans le cas des emplacements i, à l'équation suivante

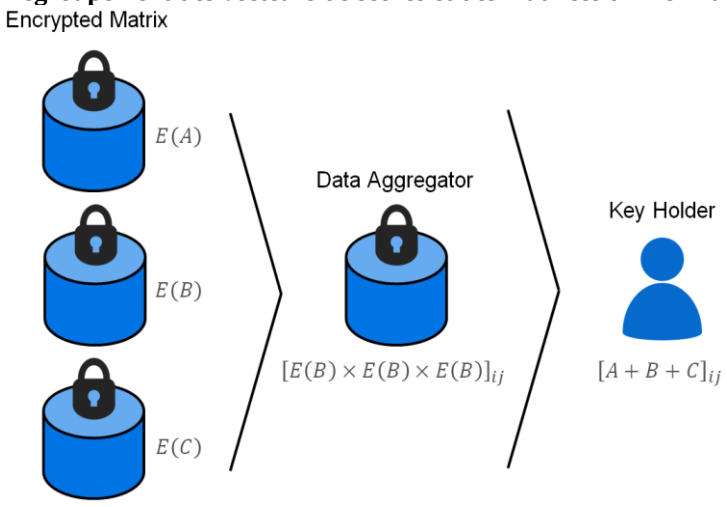
$$u(t) = \sum_i u_i(t) \quad \text{and} \quad I(t) = \sum_i I_i(t).$$

Hélas, les tentatives pour regrouper les statistiques intermédiaires exposent les emplacements à de nombreuses divulgations potentielles. Ces divulgations peuvent provenir de la matrice d'information, de la matrice de covariance, des variables indicatrices, voire des itérations elles-mêmes (pour un résumé de ces divulgations potentielles, voir l'annexe de El Emam et coll. (2012b)). L'un des principes fondamentaux du chiffrement est de prévenir absolument toutes fuites d'information; autrement, elle pourrait être utilisée pour trouver une façon d'accéder aux secrets qui sont censés être protégés.

Le calcul sécurisé peut être utilisé pour masquer tous les calculs intermédiaires. Notre protocole, appelé le Secure Pooled Analyis acRoss K-Sites (SPARK), repose sur les éléments primaires suivants : addition, multiplication, produit scalaire, multiplication de matrices, matrice inverse, distance euclidienne (norme 2) et comparaison, que nous avons pour la plupart élargis aux fins de l'application du protocole SPARK. Le protocole servant à appliquer la régression logistique distribuée sécurisée a également fait l'objet d'une évaluation visant à mesurer sa performance de calcul à l'égard d'une variété d'ensembles de données; la performance étant une préoccupation courante lorsqu'il est question de calcul sécurisé. Même en utilisant du matériel informatique générique, le temps requis pour mettre en place un modèle de régression logistique d'un million d'enregistrements dans cinq emplacements a été d'environ cinq minutes seulement (en ne tenant pas compte du temps de communication entre les emplacements).

L'exemple le plus simple d'élément primaire sécurisé utilisé dans le cadre du protocole est illustré pour l'addition de matrices à la Figure 2.3-3. Dans ce cas, nous devons simplement multiplier les messages individuels en texte chiffré qui constituent les éléments de la matrice. Bien entendu, l'application intégrale du protocole SPARK dans le cas d'une équation d'estimation généralisée (GEE) ou d'un modèle linéaire généralisé (GLM) est plus complexe que ne le laisse entendre cet exemple. Cet exemple montre néanmoins que des éléments primaires fondamentaux peuvent être dérivés des simples propriétés du système de chiffrement de Paillier et permettre de réaliser des analyses plus complexes.

Figure 2.3-3
Regroupement des vecteurs de scores et des matrices d'information à l'aide de l'addition homomorphe



3. Conclusions

Moyennant des obligations contractuelles et des mesures garantissant l'absence de fuites de renseignements personnels à partir des résultats eux-mêmes, le calcul sécurisé peut être envisagé, dans le contexte d'un cadre de travail fondé sur le risque, comme correspondant à des données pseudonymes *protégées* assorties d'un risque très faible de réidentification. Le calcul sécurisé convient bien aux scénarios caractérisés par une analyse et une collecte continues et systématiques de données, telle la surveillance en matière de santé publique, parce que les calculs peuvent être définis et optimisés, puis appliqués de façon continue.

Remerciements

Les présents travaux ont été financés par le Programme des chaires de recherche du Canada et les Instituts de recherche en santé du Canada.

Bibliographie

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, Hoboken, New Jersey.

Benaloh, J. C. (1986). "Secret Sharing Homomorphisms: Keeping Shares of a Secret Secret (Extended Abstract)." *Advances in Cryptology — CRYPTO' 86*, Lecture Notes in Computer Science, A. M. Odlyzko, ed., Springer Berlin Heidelberg, 251–260.

Centers for Disease Control and Prevention. (2004). *Integrated Guidelines for Developing Epidemiologic Profiles: HIV Prevention and Ryan White CARE Act Community Planning*.

Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine. (2015). *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. Washington (DC): National Academies Press (US);

Cousins, D., Rohloff, K., Peikert, C., and Sumorok, D. (2015). *SIPHER: Scalable Implementation of Primitives for Homomorphic EncRyption*. Final Technical Report, Raytheon BBN Technologies, Rome, NY, USA.

Dan Bogdanov, Liina Kamm, Sven Laur, and Ville Sokk. (2014). "Rmind: A Tool for Cryptographically Secure Statistical Analysis." *IACR Cryptology ePrint Archive*, 512, 1–40.

Desmedt, Y., and Frankel, Y. (1989). "Threshold Cryptosystems." *Advances in Cryptology — CRYPTO' 89 Proceedings*, Lecture Notes in Computer Science, G. Brassard, ed., Springer New York, 307–315.

Duncan, G. T., Elliot, M., and Salazar-González, J.-J. (2011). *Statistical Confidentiality*. Springer New York, New York, NY.

El Emam, K. (2013). *Guide to the De-Identification of Personal Health Information*. CRC Press (Auerbach).

El Emam, K., and Arbuckle, L. (2013). *Anonymizing Health Data: Case Studies and Methods to Get You Started*. O'Reilly.

El Emam, K., Arbuckle, L., Essex, A., Samet, S., Eze, B., Middleton, G., Buckeridge, D., Jonker, E., Moher, E., and Earle, C. (2014). "Secure Surveillance of Antimicrobial Resistant Organism Colonization or Infection in Ontario Long Term Care Homes." *PLoS ONE*, 9(4), e93285.

El Emam, K., Hu, J., Samet, S., Peyton, L., Earle, C., Jayaraman, G., Wong, T., Kantarcioglu, M., and Dankar, F. (2012a). "A Protocol for the Secure Linking of Registries for HPV Surveillance." *PLoS ONE*, 7(7).

El Emam, K., Samet, S., Arbuckle, L., Tamblyn, R., Earle, C., and Kantarcioglu, M. (2012b). "A Secure Distributed Logistic Regression Protocol for the Detection of Rare Adverse Drug Events." *Journal of the American Medical Informatics Association*.

Gueron, S., Lindel, Y., Nof, A., and Pinkas, B. (2015). "Fast Garbling of Circuits Under Standard Assumptions." *22nd ACM Conference on Computer and Communications Security*, Denver, CO, 1–43.

Health Information Trust Alliance. (2015). *HITRUST De-Identification Framework*. HITRUST Alliance.

Information Commissioner's Office. (2012). *Anonymisation: Managing Data Protection Risk Code of Practice*. Information Commissioner's Office.

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., and Walford, N. (1991). "The Case for Samples of Anonymized Records From the 1991 Census." *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 154(2), 305–340.

Morgan, M. G., Henrion, M., and Small, M. (1992). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge; New York.

Office for Civil Rights. (2012). *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. Department of Health and Human Services, Washington, DC.

O'Keefe, C., and Chipperfield, J. (2013). "A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems." 81(3), 426–455.

Paillier, P. (1999). "Public-key cryptosystems based on composite degree residuosity classes." *Proceedings of the 17th international conference on Theory and application of cryptographic techniques*, EUROCRYPT'99, Springer-Verlag, Berlin, Heidelberg, 223–238.

PhUSE De-Identification Working Group. (2015). *De-Identification Standards for CDISC SDTM 3.2*.

Rohloff, K., and Cousins, D. B. (2014). "A Scalable Implementation of Fully Homomorphic Encryption Built on NTRU." *Financial Cryptography and Data Security*, Lecture Notes in Computer Science, R. Böhme, M. Brenner, T. Moore, and M. Smith, eds., Springer Berlin Heidelberg, 221–234.

Shamir, A. (1979). "How to Share a Secret." *Commun. ACM*, 22(11), 612–613.

Songhori, E. M., Hussain, S. U., Ahmad-Reza, S., Schneider, T., and Koushanfar, F. (2015). "TinyGarble: Highly Compressed and Scalable Sequential Garbled Circuits." *2015 IEEE Symposium on Security and Privacy*, San Jose, CA, 411–428.

Statistics Canada. (2007). "Therapeutic Abortion Survey." <http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3209>.

Subcommittee on Disclosure Limitation Methodology. (2005). "Statistical Policy Working Paper 22 - Report on Statistical Disclosure Limitation Methodology." Federal Committee on Statistical Methodology.

The Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation. (2015). *Accessing Health And Health-Related Data in Canada*. Council of Canadian Academies.

Vose, D. (2008). *Risk Analysis: A Quantitative Guide*. Wiley, Chichester, England ; Hoboken, NJ.

Yao, A. C.-C. (1986). "How to Generate and Exchange Secrets." *27th Annual Symposium on Foundations of Computer Science, 1986*, Toronto, ON, 162–167.