Practical Applications of Secure Multiparty Computation for Public Health and Post-Marketing Drug Surveillance

Luk Arbuckle[1], Khaled El Emam[1,2]
[1] Children's Hospital of Eastern Ontario Research Institute, Ottawa, Canada
[2] Faculty of Medicine, University of Ottawa

Abstract

Accessing data for analytics purposes can be facilitated if the data is no longer considered personal information. Statistical and computational disclosure control methods can be used for that purpose. Another approach that is more suitable under some use cases is secure (multiparty) computation. Secure computation is well suited in the context of surveillance problems because the computations can be defined and optimized and then continuously applied. We will explain how secure computation can be applied in health surveillance contexts, with some theoretical and practical results. One project developed a secure linking protocol to link different data sets without sharing personal information. Public health use cases for secure linking include performing HPV vaccination evaluations and Chlamydia testing from general practices. Another example is antimicrobial resistant infection surveillance in long term care homes. This allowed the collection of colonization and infection data without revealing the rates for any of the participating homes. We will also discuss a post-marketing surveillance project where rare drug adverse events are modeled using logistic regression by securely pooling data from multiple sites.

Key Words: Security, privacy, risk, disclosure control, de-identification, anonymization

1. Introduction

There is growing demand to share health data for secondary purposes—purposes other than providing health care to patients—including health services research and public health. In many cases there is a need or desire to combine data from multiple sources. However, combining datasets from different data custodians or jurisdictions to perform an analysis on the pooled data creates significant privacy concerns that would need to be addressed. Legislation will not typically allow the disclosure of personal health information without consent (*Health Insurance Portability and Accountability Act* 1996, *Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data* 1995 p. 95). Consent will always be limited to a smaller population, most likely from specific socioeconomic subgroups (e.g., based on race, education) thereby biasing research (El Emam et al. 2013). Furthermore, consent cannot actually address privacy concerns in a meaningful way.

Healthcare organizations are concerned about patient privacy for a variety of reasons. The number of organizations affected by health data breaches in a given year ranges from 19 to 27% according to surveys (HIMSS Analytics 2010, 2012). The costs of a breach are significant, with healthcare being the most expensive globally at an estimated $363 per capita (Ponemon Institute 2015). Direct costs include escalation and breach notification,

hiring external experts, legal fees, and offering identify protection to affected individuals; indirect costs include internal efforts to organize the response or investigate the incident, and the loss of customers. However, there is also the loss of business opportunities due to reputational damage, and possibly litigation initiated by the affected parties.

We will consider two options for sharing health data: risk-based de-identification; and secure computation. We will further describe how the latter can be considered in a risk-based framework, before we review advanced techniques for sharing health data and summarize some of our work in this area.

1.1 Risk-Based De-Identification

The examples of AOL (Barbaro and Zeller Jr. 2006 p. 4417749) and the attack on the Netflix Prize (Narayanan and Shmatikov 2008) are often given as examples of failed attempts to protect against re-identification when data is used for other purposes, and that de-identification is therefore unable to protect data custodians from the inevitability of a successful attack. The standards in working with health data would not, however, consider the removal of unique or direct identifiers—i.e., pseudonymous data (*Health informatics. Pseudonymization* 2008)—as de-identification, which is all that was done to protect these datasets. In a systematic review, we found no examples of attacks on properly de-identified heath data where the risk was anything but "reasonable" or "very small" (El Emam et al. 2011).

The methods of de-identification should leverage the long history of statistical disclosure control methods that consider the relationship (marginal distribution) from key variables or quasi identifiers to protect against identity disclosure (Duncan et al. 2011). Furthermore, the regulations and guidance documents suggest a framework for de-identification that is risk based—i.e., that incorporates the context of the data release into the risk assessment framework—to have strong assurances that the risk is "reasonable" or "very small" (Office for Civil Rights 2012; Information Commissioner's Office 2012).

We use the term anonymization to mean both masking and de-identification (El Emam and Arbuckle 2013), but we will assume that masking unique or direct identifiers is well understood and instead focus on risk-based de-identification of the key variables or quasi identifiers. Consider the probability of re-identification given that an attacker makes an attempt as $\Pr(\text{reid} \mid \text{attempt})$ (Marsh et al. 1991). We can therefore formulate the problem as the probability of re-identification and an attack using

$$\Pr(\text{reid}, \text{attempt}) = \Pr(\text{reid} \mid \text{attempt}) \times \Pr(\text{attempt}).$$

The probability of an attacker attempting a re-identification is given by the context of the data release, using a subjective assessment of risk (Morgan et al. 1992; Vose 2008) based on expert opinion and precedent (e.g., (Centers for Disease Control and Prevention 2004; Statistics Canada 2007; Subcommittee on Disclosure Limitation Methodology 2005). The factors that affect an attempt include the security and privacy practices of the data requestor (e.g., controlling access, disclosure, retention, and disposition of data, accountability and transparency) and contractual obligations (e.g., prohibit re-identification, prohibit sharing of data without the data custodian's knowledge, and audit requirements). Furthermore, a defensible risk threshold can be determined based on precedent by evaluating the potential invasion of privacy (e.g., the sensitivity of the data, potential injury, and appropriateness of consent).

All of the above factors can be used to formulate a repeatable risk assessment framework (El Emam 2013). In fact many standards and guidelines have incorporated such a risk-based framework in their recommendations for sharing health data (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; The Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation 2015).

1.2 Secure Computation

Assume that multiple parties want to pool data and compute a function without either party knowing their inputs. The basic idea of secure computation is to compute a function on encrypted data, without ever decrypting the data to achieve the desired output. Cryptographic primitives, or building blocks, to create secure computation protocols can come from homomorphic encryption, garbled circuits, secret sharing, or others, each with their own advantages and disadvantages.

Homomorphic encryption became practical with the introduction of the Paillier cryptosystem because computation time is reasonable, but it has a limited set of operations (Paillier 1999). We will discuss this cryptosystem in more detail in the next section and in the discussion of practical applications of secure computation for public health and post-marketing drug surveillance. Yao's garbled circuits (Yao 1986) have been considered impractical due to computation time and memory requirements, although new methods may change this (Gueron et al. 2015; Songhori et al. 2015). A secret sharing scheme such as Shamir's (Shamir 1979) would be computationally efficient in a homomorphic scheme (Benaloh 1986) but parties must reveal their shares at the time of secret reconstruction (i.e., to get the final output), and therefore other techniques must be employed (Desmedt and Frankel 1989).

Secure computation is consistent with guidance provided by regulators as a means to protect personal health information while sharing encrypted data for the purposes of collaborative analysis. With secure computation the highest levels of security controls are implied, and there would be no processing of personal health information by humans. With appropriate contractual obligations and measures to ensure there are no leakages of personal information from the results themselves (O'Keefe and Chipperfield 2013), secure computation can be thought of in a risk-based framework as *protected* pseudonymous data with a very low risk of re-identification.

Secure computation can therefore alleviate one of the key barriers to the establishment of large-scale public health surveillance programs. It is also well suited to these programs because the computations can be defined and optimized and then continuously applied.

2. Homomorphic Encryption

The Pallier cryptosystem (Paillier 1999) allows us to perform operations on encrypted data, called cyphertext messages, that map to operations on the raw data, called plaintext messages. If the plaintext message is m, we will use $E(m)$ to denote the cyphertext message. This cryptosystem requires a public key and a private key, known as asymmetric cryptography. The public key is used by the parties to encrypt their data, whereas the private key, held by a semi-trusted third party, is used to decrypt the results. The basic setup is show in Figure 1.

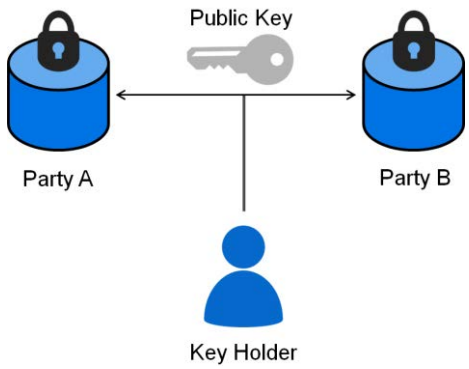

Figure 1: Basic two-party setup of a Paillier cryptosystem.

It is worth noting that the Paillier cryptosystem uses randomness in its encryption algorithm—encrypting the same message several times will yield different ciphertext messages except in rare cases. This ensures that an adversary holding a public key would not be able to compare an encrypted message to all possible encrypted counts from zero onwards and determine what the original plaintext value is.

The Paillier cryptosystem supports addition and a limited form of multiplication. For two plaintext messages a and b, plaintext addition is given by

$$E(a) \times E(b) = E(a + b),$$

where the multiplication of cyphertext messages is $\mathrm{mod}\, p^2$, and the addition of plaintext messages is $\mathrm{mod}\, p$ (for p the product of two large prime numbers). That is, multiplying the two cyphertext messages $E(a)$ and $E(b)$ is equivalent to the addition of the plaintext messages (which is encrypted). Decrypting the latter reveals the sum. Multiplication of two plaintext messages a and b, however, requires the plaintext message b (not the cyphertext message) be applied to the cyphertext message $E(a)$, because it is given by

$$E(a)^b = E(a \times b),$$

where the exponentiation of cyphertext messages is $\mathrm{mod}\, p^2$, and the multiplication of plaintext messages is $\mathrm{mod}\, p$. That is, the exponentiation of the cyphertext message $E(a)$ by the plaintext message b is equivalent to the multiplication of the plaintext messages (which is encrypted). Decrypting the latter reveals the product.

The limited form of multiplication is due to the use of a private key, but complex protocols can nonetheless be written to work around this limitation. We will see this in the examples that follow, especially the secure protocol to implement generalized linear models.

3. Secure Linking

Linking can be used to perform record lookup, or database matching for deduplication. Often, however, the best fields for linking are ones that cannot be disclosed (e.g., social security number, medical record number, health insurance number, and first and last name). The goal of secure linking is to link without sharing sensitive or personal

information, and this is not limited to just the fields used for linking. In many cases registries, for example, cannot learn new information about patients without consent or additional authorization. Revealing to another party which patients are in a registry may itself be a disclosure if membership to the registry is sensitive—membership could imply treatment for drug abuse, mental health treatment, or social assistance.

In our secure linking protocol (El Emam and Arbuckle 2013 chap. Secure Linking) we use secure computation with a semi-trusted third party (sTTP)—trusted to run the protocol, but unable to obtain sensitive or personal information about patients even if it wanted to. Because the data and the computations on the data are protected by encryption, the parties do not need to trust one another or the sTTP. None of the parties involved can "peak" into the data or the computations. Furthermore, a breach at any one site would not reveal the identity or personal information of patients.

Referring to Figure 1, the first step in our secure linking protocol is for the key holder to generate private and public keys, and to distribute the public key to the data custodians who will use it to encrypt the link variables. The encrypted data will then be shared with a central aggregator, or one data custodian sends encrypted data to the other. A homomorphic equality test is then run on the pooled encrypted data. The encrypted match results are then sent to the key holder, who uses the private key to decrypt the results.

This secure protocol is used by the Institute for Clinical Evaluative Sciences (ICES) for linking de-identified data (matching on insurance number, name, and date of birth), and was proposed to determine Chlamydia screening and testing rates with a public health agency (matching electronic medical records from family doctors to lab testing). It was also proposed for a human papillomavirus (HPV) vaccine initiative impact assessment, where more details about the protocol can be found (El Emam et al. 2012a).

4. Secure Surveillance

The collection and analysis of data for the purposes of health monitoring and intervention requires access to relevant health data. Privacy legislation and concerns can hamper these efforts, even when it is permissible to use the data for the purposes of public health. Secure protocols can, however, overcome many of these barriers by ensuring no sensitive or personal information is disclosed at the point of data collection or during the calculation of relevant statistics.

4.1 Prevalence of ARO's

Using a secure data collection system, providing strong privacy and confidentiality assurances, we were able to conduct a point prevalence study to assess rates of antimicrobial resistant organisms (ARO) in long term care homes in Ontario (El Emam et al. 2014). Although there is stigma attached to the identification of residents carrying ARO's in long term care homes, secure computation allowed the collection of colonization and infection data without revealing the rates for any of the participating homes. This addressed the need to collect data about the prevalence of ARO's in long term care homes for public health surveillance and intervention purposes.

The basic framework of the secure data collection system can be seen in Figure 2. Long term care homes provided the counts, which were encrypted at the point of collection. These encrypted counts were then combined by a data aggregator using secure computation, which then passed the intermediate results to the key holder for decryption.

Secure computation was used to determine the mean colonization or infection rates and the standard deviation, by region and facility size, and to run a two-sample randomization test (randomized t-test) for non-response bias.

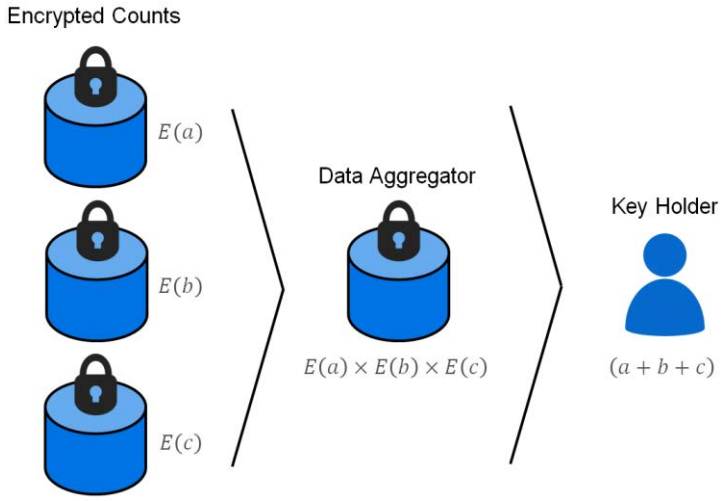

Figure 2: Computing point prevalence using a Paillier cryptosystem.

All long term care homes in the province were asked to provide colonization or infection counts for methicillin resistant Staphylococcus aureus (MRSA), vancomycin-resistant enterococci (VRE), and extended-spectrum beta-lactamase (ESBL) as recorded in their electronic medical records, and the number of current residents. We give an example of the results for MRSA only in Table 1, where the empty cells are due to minimum cell size requirements.

Table 1: MRSA cases per 100 residents, where $\boldsymbol{\rho}$ is the regional prevalence.

	Regions							
No. of beds	*North*	*East*	*Central East*	*Toronto*	*Central West*	*West*	*ρ*	*s.d.*
1-60	1.57	3.17	0.72	-	3.31	8.38	3.87	3.24
61-120	1.07	2.04	1.80	-	2.73	7.88	3.34	2.85
121-180	0.56	2.54	1.08	0.91	3.15	7.83	2.94	2.58
180+	-	2.37	1.68	2.58	2.91	8.63	2.61	2.10
ρ	0.79	2.42	1.44	1.86	3.00	8.04		
s.d.	0.46	0.38	0.42	1.15	0.22	0.37		

Data was collected online during the October-November 2011 period. Overall, 82% of the homes in the province responded, which is much higher than in previous attempts to collect data (without the use of secure computation). The microbiological findings and their distribution were consistent with available provincial laboratory data reporting test results for AROs in hospitals.

The burden of ARO in long term care settings in Ontario had not been measured at the time of our study. There is no current requirement to report ARO colonization and infection rates to the public or to public health authorities. When surveillance data is unavailable, it can be difficult to make informed decisions and to identify needs.

4.2 Rare Adverse Drug Events

Logistic regression is commonly used in the analysis of adverse drug events. Data needs to be pooled, however, to detect rare events and ensure sufficient population heterogeneity to ensure the safety and effectiveness of a drug for subpopulations. We therefore developed a secure distributed logistic regression protocol using a single analysis center with multiple sites providing data (similar to the previous example) for post-marketing surveillance. We also extended the protocol to use generalized estimating equations (GEE) to account for correlated data, other generalized linear models (GLM), and survival models (El Emam et al. 2012b).

To estimate a generalized linear model (Agresti 2002), we can use the Newton-Raphson method and iteratively compute the parameter estimates b using

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

where $u(t)$ is the score vector and $I(t)$ is the information matrix for iteration t. With multiple sites contributing data (horizontally partitioned, so that they have the same covariates and coding formats), the score vector and information matrix are in fact computed individually at each site, and combined later. That is, for i sites,

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

Unfortunately attempting to pool the intermediate statistics leaves the sites open to many potential disclosures. These disclosures can come from the information matrix, the covariance matrix, indicator variables, even the iterations themselves (for a summary of these potential disclosures see the appendix to El Emam et al. (2012b)). A core tenet of cryptography is to avoid any and all leakages of information—otherwise it could be used to find a way to extract the secrets that are meant to be protected.

Secure computation can be used to hide all of the intermediate computations. Our protocol, called Secure Pooled Analyis acRoss K-Sites (SPARK), uses the secure building blocks of addition, multiplication, dot product, matrix multiplication, matrix inverse, and two-norm distance and comparison, many of which we extended for the purposes of implementing SPARK. The protocol to implement secure distributed logistic regression was also evaluated to assess its computational performance on a variety of datasets, as performance is a common concern with the use of secure computation. Even on commodity hardware the time it took to fit a logistic regression model of one million records across five sites was only about five minutes (disregarding the communication time between sites).

The simplest example of a secure building block being used in the protocol is shown for matrix addition in Figure 3. In this case we simply need to multiply the individual cyphertext messages which are the matrix elements. Of course, the complete implementation of SPARK for a GEE or GLM is more complicated than this example would suggest. Nonetheless it shows that from the simple properties of the Paillier cryptosystem basic building blocks can be derived that allow for more complicated analysis to be performed.

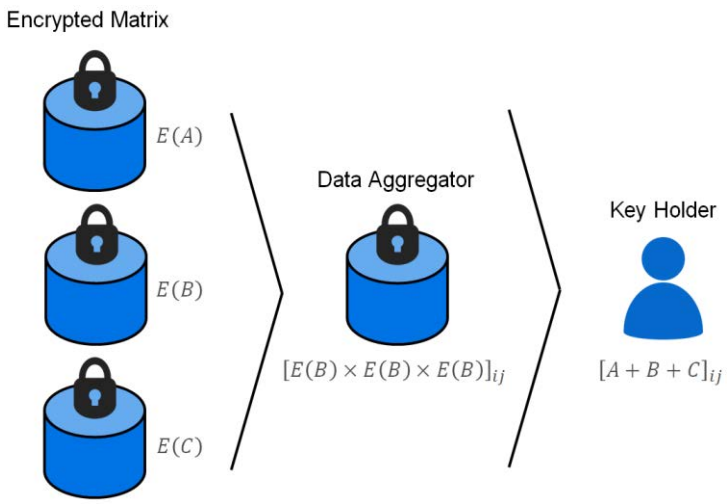

Figure 3: Pooling score vector and information matrix is simply homomorphic addition.

5. Conclusions

Secure computation—with the appropriate contractual obligations and disclosure controls on the output of computations—can be seen in a risk-based framework as *protected* pseudonymous data with a very low risk of re-identification, thereby making it consistent with regulatory guidance. It allows for the pooling of data for statistical purposes without the need to disclose personal information, and protects data custodians in the case of breach since the raw data and computations are encrypted. Examples of secure protocols for the collection, sharing, and computation of statistics exist and are practical in real settings. In particular, secure computation has been effectively demonstrated for cases of public health surveillance where there is a need to securely collect and analyze data for the purposes of health monitoring and intervention.

Acknowledgements

This work was funded by the Canada Research Chairs program, and the Canadian Institutes of Health Research.

References

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

Barbaro, M., and Zeller Jr., T. (2006). "A Face Is Exposed for AOL Searcher No. 4417749." *New York Times*.

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);

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.

Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data. (1995). .

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., Jonker, E., Arbuckle, L., and Malin, B. (2011). "A Systematic Review of Re-identification Attacks on Health Data." *PLoS ONE*, 6(12).

El Emam, K., Jonker, E., Moher, E., and Arbuckle, L. (2013). "A Review of Evidence on Consent Bias in Research." *American Journal of Bioethics*, 13(4), 42–44.

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 informatics. Pseudonymization. (2008). International Standard, ISO.

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

Health Insurance Portability and Accountability Act. (1996). *Pub. L. No. 104-191, 110 Stat. 1936.*

HIMSS Analytics. (2010). *2010 HIMSS Analytics Report: Security of Patient Data*. HIMSS.

HIMSS Analytics. (2012). *2012 HIMSS Analytics Report: Security of Patient Data*.

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.

Narayanan, A., and Shmatikov, V. (2008). "Robust De-anonymization of Large Sparse Datasets." *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, 111–125.

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*.

Ponemon Institute. (2015). *2015 Cost of Data Breach Study: Global Analysis*.

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://www.statcan.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SDDS=3209&lang=en&db=IMDB&dbg=f&adm=8&dis=2#b9>.

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.