Engineering risk-based anonymisation solutions for complex data environments

Received: 15th May, 2020

Luk Arbuckle
Chief Methodologist, Privacy Analytics, Canada

Luk Arbuckle is Chief Methodologist at Privacy Analytics, providing strategic leadership in how to responsibly share and use data. Luk was previously Director of Technology Analysis at the Office of the Privacy Commissioner (OPC) of Canada, leading a highly skilled team that conducted privacy research and assisted in investigations when there was a technology component involved. Before joining the OPC, he worked on developing methods of anonymisation and identifiability measurement tools, participated in the development and evaluation of secure computation protocols, and led a top-notch research and consulting team that developed and delivered data anonymisation solutions. He is author of two books about data anonymisation as well as numerous papers and guidance documents. Previously, Luk did both graduate and industry research in applied statistics and digital image processing and analysis.

251 Laurier Avenue W, Suite 200, Ottawa, Ontario, Canada, K1P 5J6
Tel: 001-613-368-9516; E-mail: larbuckle@privacy-analytics.com

Muhammad Oneeb Rehman Mian
Senior Data Scientist, Privacy Analytics, Canada

Muhammad Oneeb Rehman Mian is a Senior Data Scientist at Privacy Analytics, specialising in cutting-edge privacy engineering solutions. His work entails developing scalable risk-based anonymisation technologies, improving threat modelling and identifiability metrics, and exploring solutions to practical challenges faced in privacy engineering. Previously, Muhammad did graduate and postdoctoral work in biomedical research, utilising applied statistics and Big Data analytics.

251 Laurier Avenue W, Suite 200, Ottawa, Ontario, Canada, K1P 5J6
Tel: 001-613-368-4313; E-mail: mrehman@privacy-analytics.com

Abstract Technological advancements have dramatically increased the ability to collect, store and process vast quantities of data. The general applicability and precision of analytical tools in artificial intelligence and machine learning have driven organisations to leverage these advances to process personal data in new and innovative ways. As stewards of personal data, organisations need to keep that data safe and ensure processing is legal and appropriate. Having more data, however, has also led to an increased interest to process personal data for purposes other than why they were originally collected, known as secondary purposes. The reuse of personal data introduces important regulatory challenges, increasing the need to disassociate data used for secondary purposes from personal data, be it to safeguard the data, support a legitimate interest, or anonymise the data. Whereas some academics have focused on specific issues preventing more widespread adoption of this privacy-enhancing technology, others have reframed the discussion around anonymisation as risk management. Combining technology-enabled processes with measures of identifiability provides an opportunity to meet complex business needs while ensuring best practice is adopted in reusing sensitive data. This paper examines these many considerations and demonstrates how risk-based anonymisation can and should be detailed, evidence based and objectively supported through measures of identifiability. The engineering of privacy solutions, through the application of risk-based anonymisation, is also briefly explored for complex use cases involving data lakes and hub and spoke data collection, to provide the reader with a deeper understanding of real-world risk-based anonymisation in practice.

KEYWORDS: anonymisation, de-identification, pseudonymisation, privacy engineering, secondary uses

INTRODUCTION

Anonymisation can help facilitate the reuse of personal data for secondary purposes, although it is an area where privacy engineering is significantly challenged due to the intricacies and complexity of large-scale data collection and use. Nevertheless, the reuse of data is of tremendous importance to organisations trying to make the most of data, and it can provide many benefits to individuals, society and industry. It can help drive service improvements, spark innovative developments in existing or new areas, and drive a deeper and more meaningful understanding of human interactions and conditions. By asking new and innovative questions, the reuse of data can create novel insights and help find new research and development directions that can lead to better and more targeted interventions and services.

Nowhere is this clearer than in the fields of healthcare and medicine. Although already well established, at the end of 2019, the European Parliament endorsed a resolution on the digital transformation of health and care, recommending access to and sharing of personal health data while respecting strict privacy rules.[1] And in recent years, there has been a push to increase the sharing and reuse of clinical trials data. With the clinical trials transparency initiatives of the European Medicines Agency[2] and Health Canada[3] as well as industry-led initiatives such as YODA[4] or more recently Vivli,[5] clinical trials data is being made increasingly available for secondary purposes. Secondary use of clinical trials data can ease the burden on research subjects by reducing data collection requirements and making better use of subjects' contributions.[6] Other fields also benefit from the reuse of data.

The reuse of personal data for secondary purposes requires a legal basis for processing, and it is incumbent upon an organisation to ensure the purposes for reuse are appropriate and that they can demonstrate the benefits that will help justify the processing to stakeholders. A range of considerations exist depending on jurisdiction, type of data and type of processing. This can be an involved process that introduces significant regulatory burdens on organisations. In a world of Big Data, manual assessments and human intervention are challenged in keeping up with the demands of modern privacy regulations and organisational needs; from data collection to processing and third-party data sharing, such approaches have increasingly become a bottleneck to unlocking the true potential of data that an organisation possesses. Organisations are therefore moving towards operationalising privacy technology solutions for their compliance requirements.

Privacy-enhancing technologies (PETs) ensure regulatory compliance is achieved in a reliable and reproducible manner. While specific expertise may be required in complex use cases, PETs allow for standardisation of practices within an organisation. A well-designed privacy engineering solution can effectively address the needs of an organisation while fulfilling regulatory requirements. The simplification and flexibility afforded by incorporating PETs into the data life cycle can help make a system predictable, manageable and disassociated, thereby producing a trustworthy system.[7]

This paper examines the regulatory and practical considerations of anonymisation, to demonstrate how anonymisation can and should be framed as a privacy-enhancing risk management tool. A risk management framework known as the Five Safes is then described to provide a detailed, evidence-based approach to evaluate measures of identifiability in risk-based anonymisation, thereby providing a scalable and proportionate approach to compliance, resulting in solutions that ensure data is useful while being sufficiently protected. Finally, the complexities of engineering real-world applications involving data lakes and hub and spoke data collection are briefly explored to provide the reader with a deeper understanding of risk-based anonymisation in practice.

Regulatory considerations of secondary uses

Although the most well-known legal basis for processing data remains the explicit consent of the data subject, obtaining explicit consent for secondary use of data can be difficult; and in some contexts, such as research, Big Data analytics and machine learning, obtaining explicit consent may be impractical or impossible. As a result, there are provisions in the legislation that allow for the processing of personal data on a basis other than consent of the data subject. Recognising the benefits of reusing data for innovative purposes, the European General Data Protection Regulation (GDPR) as well as other regulations such as the US Health Information Portability and Accountability Act (HIPAA) allows for the secondary use of data for specific purposes such as those that are compatible with (GDPR) or incidental to (HIPAA Privacy Rule) the original processing, legal purposes, purposes in the data subjects' best interest, purposes in the public interest, public health-related purposes, and research and statistical purposes and in the pursuit of the legitimate interest of the controller (GDPR Article 6[1]).

Some research and analysis purposes may be best categorised as legitimate interests of the controller, particularly when research is the primary purpose for processing.[8] Under the GDPR, 'legitimate interests' is recognised as a legal basis for processing where the 'processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child' (GDPR Article 6[1][f]). Inherent in these criteria is a balancing test between the interests of the data controller and the rights and freedom of the data subject.

Data transformations to reduce identifiability, or disassociate the personal from data, can reduce the risks to the rights and freedom of data subjects, thereby supporting the processing based on legitimate interests. In fact, industry has recognised a spectrum of identifiability, from pseudonymisation to anonymisation.[9] Regulations such as GDPR can also be seen as incorporating a range of identifiability, with obligations commensurate to the level of identifiability.[10] Whereas the principles of data protection do not apply to anonymised data, as evidenced by GDPR Recital 26, anonymisation can be considered more holistically as contributing to the safeguarding of data by preventing the loss of personal data, and providing assurances to individuals that nonpersonal data is being used for research or product and service development.[11] It is, however, important to consider what is meant by anonymisation.

Understanding anonymisation

Too often, the academic and policy debates around anonymisation are focused on the endpoint of determining whether data is 'anonymous' and what can be considered reasonable in the context of identifiability under privacy laws, focusing on examples where it is claimed data is not anonymous.[12] It is important to evaluate vulnerabilities in data anonymisation, as this can help inform the design of technologies that reduce identifiability to a reasonable degree. Evaluating anonymisation based on the data alone, however, implies an almost certain decrease in the information that can be shared and used for secondary purposes.[13]

A common misconception in the evaluation of anonymisation is to compare it to encryption, and this demonstrates the challenge with focusing on the endpoint alone. In encryption, the aim is to prevent an eavesdropper from learning anything from an encrypted exchange of messages between sender and receiver. This model, however, makes strong assumptions that the receiver, who can fully decrypt the original message, is always a trusted recipient who will not misuse the data contained in the message. Contrast this with anonymisation, in which the receiver is also considered a form of eavesdropper who must be protected against.[14] It is therefore impossible to conceptualise anonymisation in the same way as encryption, but also unnecessary. The receiver of anonymised data is, in fact, conceptually very different from the eavesdropper in the encryption example in another important way — the receiver can be bounded in their actions based on technical and organisational controls. In other words, the context of responsible data sharing and use needs to be factored into the evaluation of anonymisation.

It is therefore important to recognise the context in which data is shared and used to understand the probability of a vulnerability being exploited in the first place.[15] Besides the receiver, other potential adversaries need to be considered to determine what is called the 'threat landscape'. This is consistent with the modelling of threat sources used in information security and risk modelling.[16] The concept of evaluating vulnerabilities and putting them in context is one that is well understood in the field of data security. It is an approach that has also been identified as critical for moving the debate forward with a more meaningful focus on the process of anonymisation and risk.[17] This can be best encapsulated by describing risk-based anonymisation.

RISK-BASED ANONYMISATION

Risk-based anonymisation is both technology and risk management. It can be thought of as reducing identifiability in support of a risk management framework to address privacy concerns — it is a process of minimising risk.[18] This approach to anonymisation has also been called process based and is highly dependent on the context of data sharing and use. The focus on process to limit the threat landscape introduces transaction costs that reduce the likelihood of a vulnerability being exploited.[19]

Identifiability in this context is still subject to a reasonableness standard, and a risk-based approach is supported by most privacy laws, such as the GDPR, in making that determination. This is reflected in industry efforts to standardise regulatory guidance through risk management frameworks for anonymisation,[20] and in emerging industry standards.[21] These efforts help support best practice with regular updates and wide adoption, creating an effective baseline for evaluating implementations.

A critical feature of a risk-based approach is that data deemed anonymised in one context may become personal data if the context changes (because of changes in purpose for processing, the intended recipients, or technical or organisational controls). Risk-based anonymisation should be thought of as a dynamic risk management process, requiring oversight and review on an ongoing basis with regular risk assessments to ensure residual risks are minimised to a reasonable degree.[22] The release-and-forget model has limited applicability outside of public data sharing, and perhaps not even in that use case as potential vulnerabilities may be identified and action needed to reduce risk exposure.[23]

The goal of risk-based anonymisation is to manage residual risks that would otherwise remain when all other factors defining the context of data sharing and use are taken into consideration. By starting with context, one minimises data transformations needed to reduce identifiability, or disassociate data from personal information, while ensuring useful data is available for innovative purposes. The factors that are included in risk-based anonymisation can be summarised using an established framework for data sharing and use, known as the Five Safes.[24] It is a flexible framework for considering the many factors that will determine identifiability, with multiple possible solutions based on the degree of impact from each Safe. The safes can be worked through in a sequential fashion, from defining the project to producing analytical outputs, and this is how it is presented next.

Safe projects

In order to define an anonymisation project, a critical first step is to understand the use case for data that will be shared and used for secondary purposes so that wants and needs are evaluated, to determine opportunities for potential data minimisation, and technical and organisational controls that will be acceptable. Mapping the flows of data is conducted to understand possible limitations based on the source and method of data collection, as well as the destination and method of processing anonymised data. This will help identify legal and ethical boundaries to data processing, and set the criteria and constraints around the anonymisation project.

Even though data will be anonymised, purposes also need to be specified, both at source and destination, as privacy laws can have restrictions based on data type and uses, and to ensure good-faith attempts are made at building stakeholder trust.[25] Once purposes and intent are well understood, many assumptions and limitations are put forward at this stage to define project scope.

Safe people

Once the project boundaries have been established, one can evaluate trust in the anticipated recipients of the anonymised data to behave according to established guidelines. This is the organisation and people who have been identified to work with the anonymised data, which limits the scope of the risk assessment. Recall that recipients of anonymised data are considered potential adversaries, with their own motives and capacity to identify individuals in data and use them for purposes that may differ from what was originally intended. Their motives can be managed, at least in part, through privacy training and contractual obligations or data sharing and use agreements. Some specific clauses have become standard practice, such as:

- Prohibiting attempts to identify or contact data subjects;
- Audit requirements to ensure agreed upon technical and organisation controls are maintained and adhered to; and
- Limits on sharing with other parties or on how and in what form that sharing can take place.

The recipient's motives may be entirely innocent and devoid of malintent, and they may yet unintentionally recognise someone in the data, depending on where they and the data subjects are from, and their circle of acquaintances (eg family, friends, coworkers).[26] This is an unavoidable byproduct of having people work with data and analytical outputs, and the reason that the usefulness of the data is always considered with respect to data protection.[27]

Safe settings

But it is insufficient to rely on the good intentions of Safe people. The reality is that the anticipated recipients will perform their analytical duties within a specified data environment, and the environment in which anonymised data is processed can have a significant impact on identifiability based on the technical and organisational controls in place. Public data sharing is the least restrictive, as there are no technical or organisational controls in place, and therefore the overall level of identifiability is high given the broad threat landscape; private data sharing is more variable but, even under normal circumstances, considered more restrictive due to enforced security and privacy practices at the data recipient's site.

These practices that decide the Safe settings determine the bounds around potential access and use, as well as the likelihood of incidents resulting in lost or stolen data. Assessments of security and privacy practices must be detailed and evidence based to provide reasonable assurances regarding the protection of anonymised data.[28] And combined, Safe people and Safe settings will determine the likelihood of an adversary exploiting a vulnerability.[29]

Safe data

Having defined the context of the data sharing and use (ie the Safe projects, Safe people and Safe settings already described), analytical measures of identifiability are used to model the clustering of data subjects within that context. For example, data can be transformed so that the identifiable features of data subjects look the same, and are therefore clustered — rather than one data subject with a unique set of identifiable features, there are multiple data subjects that share this same set of identifiable features. This implies that directly identifying information is removed or appropriately transformed, and that the clustering is on the remaining indirectly identifying features.

The clustering itself is evaluated based on context using threat modelling of plausible attacks:[30]

- Deliberate attempts to exploit vulnerabilities to identify data subjects due to a lack of sufficient controls;
- Unintentionally recognising a data subject based on knowledge of acquaintances and their identifiable information; and
- The loss or theft of data when the controls in place fail to prevent a data incident.

Threat modelling of this sort requires assumptions about the real-world, and those assumptions need to be explicit. Measuring identifiability, however, provides objective support for decision-making, so that Safe outputs can be defined.

Safe outputs

The measures of identifiability considered under Safe data are used to inform data transformation that are applied to ensure the data is nonpersonal with reasonable assurance. The degree of clustering needed is determined based on precedents from the data sharing and use of reputable public organisations, such as national statistical organisations,[31] based on subjective criteria involving:[32]

- The benefits of data processing to individuals or industry
- The sensitivity and personal nature of the data
- The potential injury to individuals from an inappropriate processing or use of the data

- The appropriateness of approval by data subjects for sharing and using the data for the intended purposes

This is also an occasion to review how the outputs of data analysis will be used. Although ethical considerations should be captured under Safe projects at the outset, it can be worth including ethical reviews to the use of outputs to ensure they align with the intended purposes originally identified at the project definition stage. Ethical uses of data have become an important topic among regulators, especially with the advances made in artificial intelligence and machine learning.[33]

REAL-WORLD ANONYMISATION SOLUTIONS

With a framework for understanding risk-based anonymisation, summarised using the Five Safes, the intricacies of engineering technology solutions can be explored. The elements captured in this framework need to be incorporated into the governance of anonymisation through technology-enabled processes. A consolidated anonymisation strategy simplifies the iterative process to achieving compliance by providing a predetermined, defensible recipe that can be applied globally to most data in a streamlined fashion. Such an approach has positive implications beyond privacy for data processing and unification efforts as well; inconsistent transformations applied to data for the purposes of compliance reduce its analytical utility. Reasonable allowance for differing contextual settings and clustering precedents between data segments become part of this overall automated recipe, thereby ensuring Safe outputs at a micro level.

Software solutions allow privacy analysts to circumvent the complexity and reduce the need for dedicated expertise in modelling identifiability of Safe data. This is apparent in the case of handling the anonymisation of complex and detailed clinical trials documents, whereby the scope and complexity of the task can be simplified through the use of automated software.[34] Natural language processing techniques are constantly improving at detecting identifying information within tabular and narrative texts of clinical trials documents. Technology-enabled solutions can better identify the transformations required to maximise data utility while producing nonpersonal data appropriate for public release. The evolving landscape of regulations and practices, however, requires process refinement on an on-going basis;[35] a challenge that can be more easily managed via automation.

While it is critical that organisations take the right actions to meet data privacy requirements, it is equally important that an auditable proof of those decisions be recorded. This serves as a benchmark for demonstrating compliance with regulations, and provides much-needed documentation of process adherence. Comprehensive implementations of risk-based anonymisation make it easier to automatically generate detailed audit trails of decisions taken to ensure compliance — detailed assessments of all aspects of the Five Safes are captured, including a history of data transformations needed to produce Safe outputs as part of the auditable proof of work completed.

Applications and challenges

In practice, the growing demands of data sharing within and across organisations are facilitated by the implementation of efficient business-to-business (B2B) integration architectures.[36] Any data privacy safeguards must therefore be compatible with B2B processes to ensure data flows remain unrestricted. An attractive motivation to pursue PETs is that a correct solution can seamlessly fit within existing data integration infrastructures.

Data lakes are widely used as central repositories of all data, structured and

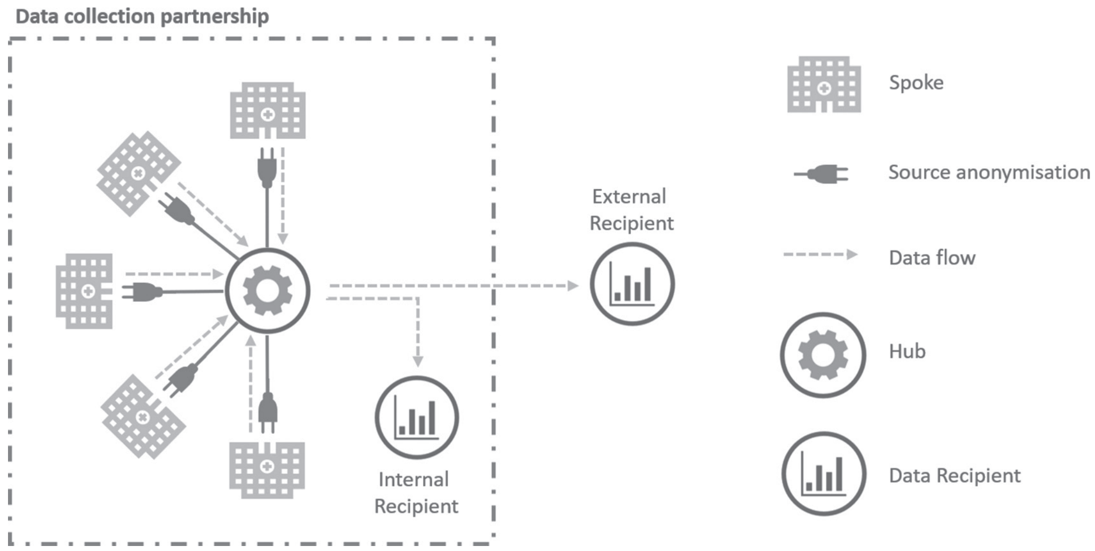

Figure 1: Source anonymisation in hub and spoke data collection

unstructured, kept in its original form typically in a distributed file system. This setup provides for a controlled, flexible and secure access to all data (as opposed to data silos) via a single data-management platform. Data lakes are particularly useful for pooling data from different sources, to gain insight from collective queries and analyses that would otherwise be not possible on single datasets. Pooling data allows small samples to be combined into a more comprehensive population, to fill in details and allow for more precise and insightful views into behaviours and impact. For example, rare events in a sample become more common in the population, meaning the rare events may no longer be disclosive.

Access to sensitive data pools by various types of stakeholders across different geo-locations, however, poses interesting privacy challenges. Threat modelling for pooled data as a whole may be underestimated when the distribution of identifiability across the pooled data is heterogeneous due to its various sources. On the other hand, disparate data formats can mean that some people appear more identifiable in the pool simply because they have fields that are not shared by other records. The security and privacy practices of recipient stakeholders may differ, which needs to be considered to ensure the safe use of data.

For these reasons, data privacy and anonymisation may need to be addressed before pooling occurs in data lakes for further reuse of data. Such complexities have made organisations cautious to 'open' their data lakes and pursue larger data-sharing opportunities. Fortunately, modern technology-enabled privacy solutions allow these issues to be taken into account without the need for dedicated data-shaping efforts to handle format and record heterogeneity within a data pool, and simplifying the application of Safe people and Safe settings in different contexts.

Cutting-edge B2B architectures circumvent some of the privacy issues highlighted earlier in other ways. In a hub and spoke model, as shown in Figure 1, data is harmonised into a single schema prior to being stored in the data hub. Unlike data lakes, this facilitates indexing and analytics. This network structure allows data from sources (the spokes) to be collected and stored in a central database (the hub), given the data being collected is often compatible or in the same format. Sources include devices or other organisations; destinations include internal or external recipients to the data collection partnership (although each data flow is assessed separately to ensure compliance). This is an opportunity to introduce privacy into design by deploying the right software-based solutions at the spokes, so that data arriving at the hub is already safe and ready to be processed or shared.

A challenge with this approach, however, is that it becomes necessary to incrementally anonymise a small number of new records being collected at the spokes without access to the rest of the data hub; new records being collected therefore appear more identifiable than they are. In this case, technological solutions can propagate a defined framework from the data hub to inform the spokes as to how new records can be clustered appropriately.[37]

When designing a privacy solution, consideration should be given to ensuring robust regulation and quality control steps are put in place to enforce best practices in the use of PETs. Deployed anonymisation technologies should also to be vetted against established codes of practice[38] and tested through deliberate attacks (such as motivated intruder tests)[39] on data deemed safe within the context of a data pipeline. Even after initial deployment, shifting subject populations and changing context settings over time mean that updated threat modelling is needed.

CONCLUSIONS

While technology-enabled privacy tools are an attractive prospect for organisations to streamline and scale their processes, it must be appreciated that these tools need to be part of a larger privacy framework with its own checks and balances. Risk-based anonymisation combines technology with risk management to ensure best practice in the reuse of personal data, to demonstrate to stakeholders that privacy is part of design. Automation streamlined processes can ensure even complex risk management-enabled anonymisation can be effectively deployed in complex use cases, meeting the needs of B2B integration architectures. Best practice in anonymisation will only achieve widespread adoption if business needs are met, demonstrating the value of privacy in practice. Risk-based anonymisation, when detailed and evidence based, meets the needs of privacy regulations while respecting their intent in data protection and privacy.

References

1. Committee on the Environment, Public Health and Food Safety (2019) 'On enabling the digital transformation of health and care in the digital single market; empowering citizens and building a healthier society', Motion for a Resolution, European Parliament, available at: https://www.europarl.europa.eu/doceo/document/B-9-2019-0239_EN.html (accessed January 2020).
2. EMA (2014) 'European Medicines Agency agrees policy on publication of clinical trial data with more user-friendly amendments', European Medicines Agency Press Office, Amsterdam, 12th June, p. 2.
3. Health Canada (2019) 'Public release of clinical information, guidance document' [online], available at: https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance/document.html (accessed January 2020).
4. Ross, J.S., Waldstreicher, J., Bamford, S., Berlin, J.A., Childers, K., Desai, N.R., Gamble, G., Gross, C.P., Kuntz, R., Lehman, R., Lins, P., Morris, S.A., Ritchie, J.D. and Krumholz, H.M.. (2018) 'Overview and experience of the YODA project with clinical trial data sharing after 5 years', *Scientific Data*, Vol. 5, No. 1, pp. 1–14.
5. Kaiser, J. (2018) 'A new portal for patient data', *Science*, Vol. 361, No. 6399, pp. 212–212.
6. Howe, N., Giles, E., Newbury-Birch, D. and McColl, E. (2018) 'Systematic review of participants' attitudes towards data sharing: a thematic synthesis', *Journal of Health Services Research & Policy*, Vol. 23, No. 2, pp. 123–133.
7. Brooks, S., Garcia, M., Lefkovitz, N., Lightman, S. and Nadeau, E. (2017) 'An introduction to privacy engineering and risk management in federal systems', National Institute of Standards and Technology, Gaithersburg, Maryland.
8. Maldoff, G. (2019) 'How GDPR changes the rules for research' [online], available at: https://iapp.org/news/a/how-gdpr-changes-the-rules-for-research/ (accessed January 2020).
9. Polonetsky, J., Tene, O. and Finch, K. (2016) 'Shades of gray: seeing the full spectrum of practical data de-

identification', *Santa Clara Law Review*,Vol. 56, No. 3, pp. 593–629.

10. Hintze, M. and El Emam, K. (2018) 'Comparing the benefits of pseudonymisation and anonymisation under the GDPR', *Journal of Data Protection and Privacy*,Vol. 2, No. 1, pp. 145–158.
11. ICO (2017) 'Big data, artificial intelligence, machine learning and data protection, guidance', Information Commissioner's Office, Cheshire, England.
12. Ohm, P. (n.d.) 'Broken promises of privacy: responding to the surprising failure of anonymization', *UCLA Law Review*,Vol. 57, p. 1701.
13. Brickell, J. and Shmatikov,V. (2008) 'The cost of privacy: Destruction of data-mining utility in anonymized data publishing', in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, Association for Computing Machinery, Las Vegas, NV, pp. 70–78.
14. Elliot, M. and Dale, A. (1999) 'Scenarios of attack: the data intruders perspective on statistical disclosure risk', *Netherlands Official Statistics*,Vol. 14, No. Spring, pp. 6–10.
15. Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J.N., Lievesley, D. and Walford, N.S. (1991) 'The case for samples of anonymized records from the 1991 census', *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, Vol. 154, No. 2, pp. 305–340.
16. ISO/IEC (2018) 'Information security risk management', International Standard, International Organization for Standardization, Berlin, Germany.
17. Hartzog, W. and Rubinstein, I. (2017) 'The anonymization debate should be about risk, not perfection', *Communication of the ACM*,Vol. 60, No. 5, pp. 22–24.
18. Garfinkel, S. (2015) 'De-identification of personal information', National Institute of Standards and Technology, Gaithersburg, MD.
19. Rubinstein, I. and Hartzog, W. (2016) 'Anonymization and risk', *Washington Law Review*, Vol. 91, pp. 703–760.
20. HITRUST Alliance (2015) 'HITRUST de-identification framework', HITRUST Alliance, Frisco, Texas.
21. ISO/IEC (2018) 'Privacy enhancing data de-identification terminology and classification of techniques', International Standard, International Organization for Standardization, Madison, Wisconsin.
22. Stalla-Bourdillon, S. and Knight, A. (2016) 'Anonymous data v. personal data - false debate: an EU perspective on anonymization, pseudonymization and personal data', *Wisconsin International Law Journal*, Vol. 34, p. 284.
23. OAIC (2018) 'MBS/PBS data publication - Commissioner initiated investigation report', Office of the Australian Information Commissioner, Sidney, Australia.
24. Arbuckle, L. and Ritchie, F. (2019) 'The five safes of risk-based anonymization', *IEEE Security & Privacy*, Vol. 17, No. 5, pp. 84–89.
25. Richards, N. and Hartzog, W. (2016) 'Taking trust seriously in privacy law', *Stanford Technology Law Review*,Vol. 19, pp. 431–472.
26. Duncan, G.T., Elliot, M. and Salazar-González, J.-J. (2011) 'Statistical confidentiality', Springer, New York, NY [online], available at: http://link.springer.com/10.1007/978-1-4419-7802-8 (accessed 10th August, 2015).
27. Duncan, G.T., Keller-McNulty, S.A. and Stokes, S.L. (2001) 'Disclosure risk vs. data uility: The R-U confidentiality map', Technical report, National Institute of Statistical Sciences, Research Triangle Park, North Carolina.
28. HITRUST Alliance (2019) 'HITRUST common security framework', HITRUST Alliance, Frisco, Texas.
29. Morgan, M. Granger, Henrion, Max and Small, Mitchell (1992) 'Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis', reprint edn, Cambridge University Press, Cambridge and New York.
30. El Emam, K. and Arbuckle, L. (2013) 'Anonymizing health data: Case studies and methods to get you started', O'Reilly, Sebastopol, California.
31. Subcommittee on Disclosure Limitation Methodology (2005) 'Statistical policy working paper 22 - Report on statistical disclosure limitation methodology' [online], availableat: http://www.fcsm.gov/working-papers/SPWP22_rev.pdf (accessed 28th October, 2011).
32. El Emam, K. (2013) 'Guide to the de-identification of personal health information', CRC Press (Auerbach), Boca Raton, FL.
33. ICDPPC (2018) 'Declaration on ethics and data protection in artificial intelligence', International Conference of Data Protection & Privacy Commissioners, Brussels, Belgium.
34. Raskind, J. (2019) 'A primer on anonymisation', *Medical Writing*,Vol. 28, pp. 44–49.
35. Keerie, C., Tuck, C., Milne, G., Eldridge, S., Wright, N. and Lewis, S. C. (2018) 'Data sharing in clinical trials – practical guidance on anonymising trial datasets', *Trials*,Vol. 19, No. 1, p. 25.
36. Bussler, C. (2003) 'B2B integration: Concepts and architecture', Springer-Verlag, Berlin Heidelberg.
37. Arbuckle, L. and El Emam, K. (2020) 'Building an anonymization pipeline: Creating safe data', O'Reilly Media, Sebastopol, California.
38. Information Commissioner's Office (2012) 'Anonymisation: Managing data protection risk code of practice', Information Commissioner's Office, Cheshire, England.
39. Good Research (n.d.) 'Motivated intruder tests'. *Good Research* [online], available at: http://motivatedintruder.com/ (accessed 29th January, 2020).