Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

doi:10.3233/978-1-61499-753-5-313

Preventing Unintended Disclosure of Personally Identifiable Data Following Anonymisation

Chris SMITH [a,1]
[a] *Leeds Institute of Health Sciences, University of Leeds, United Kingdom*

Abstract. Errors and anomalies during the capture and processing of health data have the potential to place personally identifiable values into attributes of a dataset that are expected to contain non-identifiable values. Anonymisation focuses on those attributes that have been judged to enable identification of individuals. Attributes that are judged to contain non-identifiable values are not considered, but may be included in datasets that are shared by organisations. Consequently, organisations are at risk of sharing datasets that unintendedly disclose personally identifiable values through these attributes. This would have ethical and legal implications for organisations and privacy implications for individuals whose personally identifiable values are disclosed. In this paper, we formulate the problem of unintended disclosure following anonymisation, describe the necessary steps to address this problem, and discuss some key challenges to applying these steps in practice.

Keywords. anonymisation, unintended disclosure, personally identifiable data, human judgement, privacy

1. Introduction

Personally identifiable data [14] is captured and processed by care providers to inform service provision (e.g. emergency treatment). Simultaneously, these organisations are increasingly incentivised or mandated to share datasets with other organisations, or to make data available publicly, for secondary uses (e.g. clinical research) [15].

Legislation and ethical guidance governs data sharing by organisations. Generic legislation such as the Data Protection Act [19] and Human Rights Act [20] applies in the UK. Specific legislation also exists for the health domain, including the National Health Service Act 2006 [21] and Health and Social Care Act [22] in the UK, and the Health Information Portability Act (HIPAA) [23] in the US. Ethical guidance from advisory groups such as Research Ethics Committees [9] may supplement legislation.

Anonymisation can be applied to datasets prior to sharing in order to comply with legislation and ethical guidance. Attributes in a dataset are classified by the extent to which they facilitate identification of a ``data subject'' [11]. Risk of identification [6] posed by values in specific sets of attributes is then quantified using methods such as k-anonymity [18], l-diversity [13] and β-likeness [4]. Values are transformed [8, 17] to reduce this risk below an acceptable threshold [10] for a specific context.

[1] Corresponding author, Chris Smith, Leeds Institute of Health Sciences, University of Leeds, Level 10, Worsley Building, Clarendon Way, Leeds, LS2 9NL; E-mail: c.j.smith@leeds.ac.uk.

Human judgements are used to classify attributes as: *direct identifiers*: judged to enable identification in isolation, *indirect identifiers*: judged to enable identification in conjunction with other attributes within or outside the dataset, or *non-identifiers*: judged to provide minimal risk of identification [14]. Attribute classification, along with considerations of computational tractability [3, 7] and data utility [2, 16], determine whether specific attributes are included in: i) the anonymisation process, and ii) the shared dataset.

Classification requires knowledge of how attribute values contribute to identification. Legal and ethical guidance documents [10, 23] provide pre-defined classifications of common attributes (e.g. *Surname*) to assist organisations. However, datasets to be shared may exhibit characteristics such as high-dimensionality [1], structural dynamism, and complex provenance. Consequently, classification decisions may be based on incomplete and/or imperfect knowledge regarding attributes and their values. Failure of human judgements when making these type of information security decisions [5, 25], particularly in the evaluation of risk and uncertainty [24], and the consequences in terms of security and privacy breaches [12] has been previously recognised.

We focus on a problem that may arise from errors and anomalies during the capture and processing of health data: *personally identifiable values being placed into attributes of a dataset whose values are expected to contain non-identifiable data*. For example, a unique patient identifier residing in an attribute that relates to the reason for a referral due to an erroneous processing step that transposes values between attributes, or an informal staff policy to increase the speed of internal processes. Such attributes may be judged to contain non-identifiable values and omitted from anonymisation, but then included in a shared dataset. Consequently, organisations are at risk of sharing datasets that unintendedly disclose personally identifiable values through these attributes.

2. Unintended Disclosure of Personally Identifiable Data

We model a dataset as a set of entries, $d \in D$, where each entry is a tuple, $(v_1, \dots, v_n)$, that is composed of values for a set of attributes, $\mathrm{A} = \{\mathrm{a}_1, \dots, \mathrm{a}_\mathrm{n}\}$, that relate to an individual, and V_a represents the set of distinct values held by an attribute, $a \in A$.

Patient Number	Date	Source	Reason
1120000	2014-04-01	ABC123	Cancer

Figure 1. Example entry from a dataset relating to patient referrals

Attributes are classified into one of the following distinct subsets of A to drive the anonymisation process: direct identifiers (I), indirect identifiers (Q), and non-identifiers (N), such that $\mathrm{I} \cup \mathrm{Q} \cup \mathrm{N} = \mathrm{A}$. Direct identifiers, $\mathrm{a} \in \mathrm{I}$, are removed. Non-identifiers, $a \in N$, are not considered in the anonymisation process, but may included in a shared dataset. We assume that at least one attribute is classified as a non-identifier, such that $\#(\mathrm{N}) \geq 1$. Indirect identifiers, $\mathrm{a} \in \mathrm{Q}$, are considered in anonymisation. For simplicity, and without loss of generality, we consider anonymisation to be applied over all attributes in Q rather than a subset of these attributes. Risk of identification is quantified based on combinations of values for the indirect identifiers, $V_{a_1}^* \times \dots \times V_{a_m}^*, \forall a_i \in Q$ for each entry in the dataset.

Transformations are applied to values of indirect identifiers to reduce the risk of identification. Specific transformations are dependent on the syntax and semantics of

attribute values. We do not consider aggregation over entries as a transformation, such that each entry remains associated with a single individual following anonymisation. Organisations iterate over a process of risk quantification and transformation to produce a dataset where the: i) risk of identification is reduced below an acceptable threshold, ii) data utility is sufficient for the intended use(s).

PID	Date	Source	Reason
1	2014-04	Specialist Clinic	Cancer

Figure 2. Example anonymised entry from a dataset relating to patient referrals

Figure 2 illustrates how the entry in Figure 1 might be anonymised. *Patient No* has been classified as a direct identifier and replaced by a unique (non-personal) identifier: *PID*. *Reason* has been classified as a non-identifier and retained in the dataset without transformation. *Date* and *Source* have been classified as indirect identifiers and transformed using generalisation.

Unintended disclosure of personally identifiable data has the potential to arise from judgements regarding attribute classification that are based on an expectation of the set of values, $\mathrm{E}(\mathrm{V_a})$, held by the attribute, a, rather than the actual set of values, $\mathrm{V_a}$. For example, the expectation that *Reason* will contain values that only relate to the condition for which they have been referred. Inconsistency between the expected and actual values can lead to an attribute that poses an identification risk being classified as a non-identifier and omitted from the anonymisation process. Actual values that contain personally identifiable data, such as *Patient No*, may reside in attributes classified as non-identifiers.

PID	Date	Source	Reason
1	2014-04	Specialist Clinic	Cancer-1120000

Figure 3. Example anonymised entry from a dataset relating to patient referrals with unintended disclosure

Figure 3 illustrates the problem of unintended disclosure. *Patient No* been included in the *Reason* attribute. Due to the classification of the *Reason* attribute as a non-identifier - based on the expectation that any values held by the attribute posed a minimal risk to identification - the attribute has been included in the dataset without transformation but clearly poses an identification risk.

Validation at different processing stages using techniques such as regular expressions may fail to prevent such scenarios. Methods may not be sufficiently restrictive due to their focus on the syntax rather than semantics of values. Additionally, datasets may be composed of entries from different organisations, which are subject to heterogeneous policies regarding data quality. Validation of these aggregated datasets might be insufficiently restrictive due to assumptions about upstream policies, or due to additional constraints, such as computational tractability.

3.Preventing Unintended Disclosure of Personally Identifiable Data

Prevention of unintended disclosure requires attributes to be classified based on *verification* rather than *expectation* of values. Verification ensures that any value, $\mathrm{v} \in \mathrm{V_a}$, of any attribute, $\mathrm{a} \in \mathrm{A}$, within any entry is drawn from a pre-defined set of values, V_a^*, for which the absence of personally identifiable data can be demonstrated. Any entry in the dataset would then be drawn from a defined space: $V_{a_1}^* \times \ldots \times V_{a_n}^*, \forall \mathrm{a_i} \in \mathrm{A}$. Verification could be integrated into attribute classification as follows:

- For $a \in A$, a set of values would be defined, V_a^*, which are drawn from a vocabulary for which the semantics and implications for identification of individuals are known, e.g. $V^*_{\text{Reason}} = \{\text{reason: cancer}, \text{reason: asthma}, \ldots\}$.
- For $a \in A$, the set of actual values held by the attribute, V_a, across all tuples would be determined, e.g. $V_{\text{Reason}} = \{\text{"Cancer"}, \text{"Asthma"}, \ldots\}$.
- For $v \in V_a \cap V_a^*$, a transformation function, δ, would be defined, to map actual values, V_a, of an attribute to values in V^*_a. Values not mapped would be omitted, or replaced with a *null* value, e.g. $\delta(\text{"Cancer"}) \rightarrow \text{reason: cancer}$.
- Classification would partition the set of attributes, A, based on known semantics of the extent to which they enable the identification of individuals, e.g. $I = \{\text{Patient No}\}$, $Q = \{\text{Date}, \text{Source}\}$, and $N = \{\text{Reason}\}$.

Verification would not only assist in preventing unintended disclosure, it would also provide a robust basis on which the identification risk posed by the values of different attributes could be quantified. Computational tractability and data utility could still be retained by evaluating identification risk over a subset of the attributes, Q. However, verification would ensure that classification of an attribute as a non-identifier does not risk the disclosure of personally identifiable data.

Classification based on verification of their actual values rather than human judgement is particularly important given the high-dimensionality, structural dynamism and complex provenance of datasets now captured by organisations. Robust judgement in the presence of such factors is a significant challenge for humans, yet such judgements are likely to be required more frequently within organisations in the future.

4.Challenges

Prevention of unintended disclosure presents challenges in practice, which include:

- **Computational Overhead:** Verification of attribute values against pre-defined sets is computationally intensive - requiring potentially vast numbers of comparisons. Efficient methods are required to minimise the time and resources required
- **Structural Dynamism:** Attributes and sets of attribute values can be subject to change over time - requiring changes to the verification process to ensure that it remains effective.
- **Vocabularies:** Attribute values must be compared against a pre-defined set of values - requiring relevant vocabularies to exist for each attribute. Organisations may be required to author such vocabularies if an appropriate vocabulary does not pre-exist for a particular attribute.
- **Technical Expertise:** Mapping of actual values, V_a, to a pre-defined set of values from a specific vocabulary, V^*_a, may not not necessarily be one-to-one and processing of certain formats for values may not be readily automated - requiring human involvement and technical expertise.

Without effective and efficient solutions to these challenges, organisations must decide whether to: (1) avoid sharing of datasets, or (2) share anonymised datasets and acknowledge the risk of unintended disclosure. This decision would be largely influenced by the legislative and ethical frameworks to which the organisation is subject.

5.Conclusion

Unintended disclosure of personally identifiable data has regulatory implications and poses governance challenges for data controllers. To situate sharing on a sound legal and ethical foundation, work is required to address the challenges above through novel tools and methods, vocabularies and ontologies, and education regarding anonymisation.

References

[1] Charu C Aggarwal. On K-anonymity and the Curse of Dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases*, pages 901–909, 2005.

[2] Justin Brickell and Vitaly Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 70–78, 2008.

[3] Jiwon Byun, Yonglak Sohn, Elisa Bertino, and Ninghui Li. Secure anonymization for incremental datasets. *Secure Data Management*, 4165:48–63, 2006.

[4] Jianneng Cao and Panagiotis Karras. Publishing Microdata with a Robust Privacy Guarantee. *Proceedings of the VLDB Endowment*, 5(11):1388–1399, 2012.

[5] Lorrie Faith Cranor. A framework for reasoning about the human in the loop. In *Proceedings of the 1st Conference on Usability, Psychology, and Security (UPSEC'08)*, pages 1–15, 2008.

[6] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. A systematic review of re-identification attacks on health data. *PLoS ONE*, 6(12), 2011.

[7] Gabriel Ghinita, Yufei Tao, and Panos Kalnis. On the anonymization of sparse high-dimensional data. In *Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE'08)*, pages 715–724, 2008.

[8] Yeye He and Jeffrey F Naughton. Anonymization of set-valued data via top-down, local generalization. *Proceedings of the VLDB Endowment*, 2(1):934–945, 2009.

[9] Health Research Authority. Research Ethics Committees (RECs), 2015.

[10] Information Commissioner's Office. Anonymization: Managing data protection risk - Code of Practice, 2012.

[11] ISO/TC215 Secretariat. Health Informatics - Pseudonymization (ISO/TS 25237:2008). Technical report, British Standards, 2008.

[12] Divakaran Liginlal, Inkook Sim, and Lara Khansa. How significant is human error as a cause of privacy breaches? An empirical study and a framework for error management. *Computers and Security*, 28(3-4):215–228, 2009.

[13] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-Diversity: Privacy Beyond K-Anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):1–52, 2007.

[14] Arvind Narayanan and Vitaly Shmatikov. Myths and fallacies of "personally identifiable information". *Communications of the ACM*, 53(6):24, 2010.

[15] Nuffield Council on Bioethics. The collection, linking and use of data in biomedical research and healthcare: ethical issues, 2015.

[16] Paul Ohm. Broken Promises of Privacy: Responding to the Surprising Failure of Anoymization. *UCLA Law Review*, 57:1701–1777, 2010.

[17] Pierangela Samarati and Latanya Sweeney. Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression. Technical report, SRI Computer Science Laboratory, Palo Alto, CA, 1998.

[18] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[19] UK Legislation. Data Protection Act, 1998.

[20] UK Legislation. Human Rights Act 1998, 1998.

[21] UK Legislation. National Health Service Act 2006, 2006.

[22] UK Legislation. Health and Social Care Act 2012, 2012.

[23] U.S. Department of Health and Human Services. The HIPAA Privacy Rule, 2000.

[24] Ryan West. The Psychology of Security: Why do good users make bad decisions. *Communications of the ACM*, 51(4):50–79, 2008.

[25] Charles Cresson Wood and William W. Banks. Human error: an overlooked but significant information security problem. *Computers & Security*, 12(1):51–60, 1993.