Privacy of Unstructured Data

Table of Contents

Introduction

Background

The NHS collects and stores vast quantities of structured and unstructured patient data which, when utilised downstream in research and analytics, provides insights to support evidence-based decision making and helps move towards the goal of a more efficient healthcare system.

Ensuring the correct level of patient privacy throughout this process is essential as the implications of patient re-identification with unauthorised personnel breaches a patient's confidentiality, erodes the public trust, and can place significant costs to the organisation responsible due to fines and reputational damage.

Protecting the privacy of patient data whilst maximising its utility is an ongoing research topic due to the complexity of the process, the high risks involved, and the ever growing access to larger data sources. There are numerous research and focus groups along with internal organisation solutions and commercial products which are mentioned within this report that focus efforts in balancing risk and utility.

Within this project for NHS England's Transformation Directorate (NHSE) we are focussing on the anonymisation of unstructured patient data in the form of typed or written text, for example medical notes, rather than other unstructured data e.g. radiology images, streaming data, or audio files. Unstructured data, as opposed to structured data, increases the complexity of the issue as it is more difficult to store, search and analyse and therefore is an ongoing problem yet to be solved with a robust solution.

Some examples of unstructured text in the NHS are:

- Medical Notes (discharge summary, case management, social work, pharmacy, radiology, etc.)
- Patient and Staff Feedback
- Survey Responses
- Research Papers and Clinical Trial Reports

When a person cannot be re-identified, the data is no longer considered to be 'personal data' and the GDPR does not apply to its further usage. The benefits of a robust de-identification process would be;

- To mitigate risk and minimise the harm caused to individuals from a potential data breach.
- To build community trust in how agencies store and handle data.
- To enable a Trust to safely share their data, be that internally, with another Trust or externally with third parties for the purposes of maximising insights and understanding the data.

The aim of de-identification is to reduce the risk of re-identification while retaining as much data utility as possible.

The aim of this Digital Analysis and Research Team (DART) commissioned project is for NHSE to gain a greater understanding of the considerations and issues surrounding the process of de-identification of unstructured data. In addition, it aims to identify the basic principles and the next steps for the DART towards developing a robust holistic approach to the maintenance of the privacy of unstructured data.

To aid in the production of this report, a workshop day was held on 2nd March 2022, to which significant stakeholders identified through research were invited to discuss and work through some of the underlying topics covered in this report.

Problem statement

Unstructured data (e.g. text) makes up a significant quantity of NHS data but is comparatively underused as an evidence source for analysis. This is often due to the privacy concerns restricting the sharing and use of these data.

To "our"[NHSE DART] knowledge there are currently no tools on the market that allows the NHS to robustly ascertain the level of privacy of unstructured data.

To have confidence when commissioning tooling for anonymisation purposes the NHS needs an understanding of what private content, health related text data can contain. The tooling put in place to protect the privacy of these data needs to be able to assess the content, evaluate the

risk associated with the content, and demonstrate that the tooling functionality has dealt with any privacy concerns appropriately.

Before we can seek to "solve" these issues we need to understand the wide-ranging nature of the problem. This report will aim to highlight the current topics of interest in the field, collate current examples of success and failure and bring it together to suggest a minimum criteria that needs to be considered in order to give the required level of confidence.

Report structure

This report should be used by stakeholders to understand the issues associated with using and sharing unstructured text data in a healthcare context and to start considering what tools and solutions would enable a user to address these issues robustly.

After an initial section on **Sharing Data** to further highlight the need for this work, the report presents four sections focusing on Anonymising including: the **Anonymisation Process** to discuss the wider framework the technical aspect sits within; **Assessing the re-identification risk** to discuss how to calculate the risk; a definition of **Anonymisation Levels** to support the risk setting; and some examples of current **Anonymisation Tools.**

The second half of the report starts to list out the **Challenges to anonymisation** and a series of **Case Studies and Examples** for both successful and unsuccessful implementations of dealing with privacy in unstructured data.

The key output of the report is a **List of Key Qualities for a Tool** that could be applied when considered a specific anonymisation task in order to assess a tools ability to cover the range of privacy issues highlighted in the rest of the report. These qualities link back to the rest of the content in the report.

Finally, there are some **Recommendations for Future Work** which include highlighting the need for continued knowledge share as well as continued research into appropriate tooling.

Sharing Data

Reasons to share data

It is important to highlight the significant value of data sharing. The following paper, Systematic Review effectively illustrates the benefits of data sharing, thus the requirement for anonymisation. Benefits include;

1. Health-care quality or services improvement
2. Observational risk factor-outcome research
3. Drug prescribing safety
4. Case-finding for clinical trials
5. Development of clinical decision support.

Within the review, five papers were highlighted which compared study quality with and without free text and found an improvement of accuracy when free text was included in analytical models. Another three papers using mental health data reported that more data on variables of interest were available in the free text as compared to structured data, and used free text mentions of diagnoses to augment case finding, but did not quantify the additional patient numbers found by this method.

The previously stated research paper Systematic Review has been used in the South London and the Maudsley (SLaM) model. SLaM uses a participatory governance model with a service user and career advisory group, which advises researchers on the development of their studies and on requesting linkages. This format ensures that research conducted using mental health datasets is directed by priorities identified by service users and ensures that there is a route by which service users can find out about, and become involved in the research. This reduces the separation between service users and researchers, helping to increase trust. This model of transparency in the use of patient data was favoured by members of the Brighton citizens' jury who were asked about their views on sharing their medical free text for research.

The three themes below lay out where clinical free text contributes to evidence-based research. It has proven to improve services appropriate to patients' needs, understanding who is at risk of adverse outcomes, and how drugs can be used to best treat symptoms and prevent adverse outcomes.Systematic Review

- Improving healthcare quality and service
- Understanding risk factors for disease and disease outcomes
- Improving the safety of drug prescribing.

Free text data could be used to develop automated clinical decision support, helping clinicians make decisions on patient diagnosis or treatment on the basis of a range of information in their records.[Systematic Review]

- Case identification methods
- Automated clinical decision support

Barriers to sharing data

The global pandemic has increased awareness of the vital role data plays in public policy and decision making. The general population has now grown used to daily figures and trends being shared at local and international levels, and the development of the vaccine was no doubt accelerated by increased openness between researchers and healthcare providers. In more normal times, competing researchers may be less motivated to share results or access to source data and data owners may revert to their more conservative, risk-averse habits.
The following is a list of barriers which make the sharing of data more difficult:

1. Complexity of de-identifying or anonymising data to a suitable level
2. Technical systems and data formatting that allow multiple users to access the data whilst maintaining security
3. Process time burden for checking the data to be shared and ensuring the information governance is met in line with the original data collection
4. Lack of clearly defined purpose or strong enough argument to wage the ‘risk of sharing’ against
5. User perspectives on what the data is to be used for or lack of trust in data organisations to maintain patient privacy due to lack of knowledge, or alternatively awareness of previous data breaches
6. Users’ privacy perception: different users have varying levels of concern and perceptions of privacy, driving a tendency towards the minimum risk approach

7. Context dependence: in any given situation, a person can be indifferent to their privacy while the same person may become very concerned about their right to privacy in another situation
8. Privacy paradox: users exhibit a complex and paradoxical dichotomy between their privacy (declarations) concerns and actual behaviour

Concerns around data sharing

Data specificity and dataset only understood by the data creator - One of the key arguments against the re-use of qualitative data is that the material is culturally constructed, co-produced by the researcher and researcher participants, and cannot be properly understood outside its original purpose, context, conceptual and empirical framing. A related concern from researchers goes to the heart of who owns the (often enriched) data and the ownership of knowledge built from that data; there can be a sense in which the original researcher feels they 'own' the data, or it belongs to the participants, and some reluctance for it to be appropriated by others or analysed in other ways. **Doing Research Differently**

Risk severity recognition - Perception of low level data security is treated just as severely as a real 'ethical hack' success (an actual breach being conducted by penetration testers engaged to test the system). Reputational damage is just as harmful as real-time penetration of data, e.g. DEFRA Breach report.

Data access complexity - The difficulty and complexity in data access can encourage data users to keep their own isolated repository which is therefore not governed or controlled centrally with a standardised privacy process and can be open to data breaches.

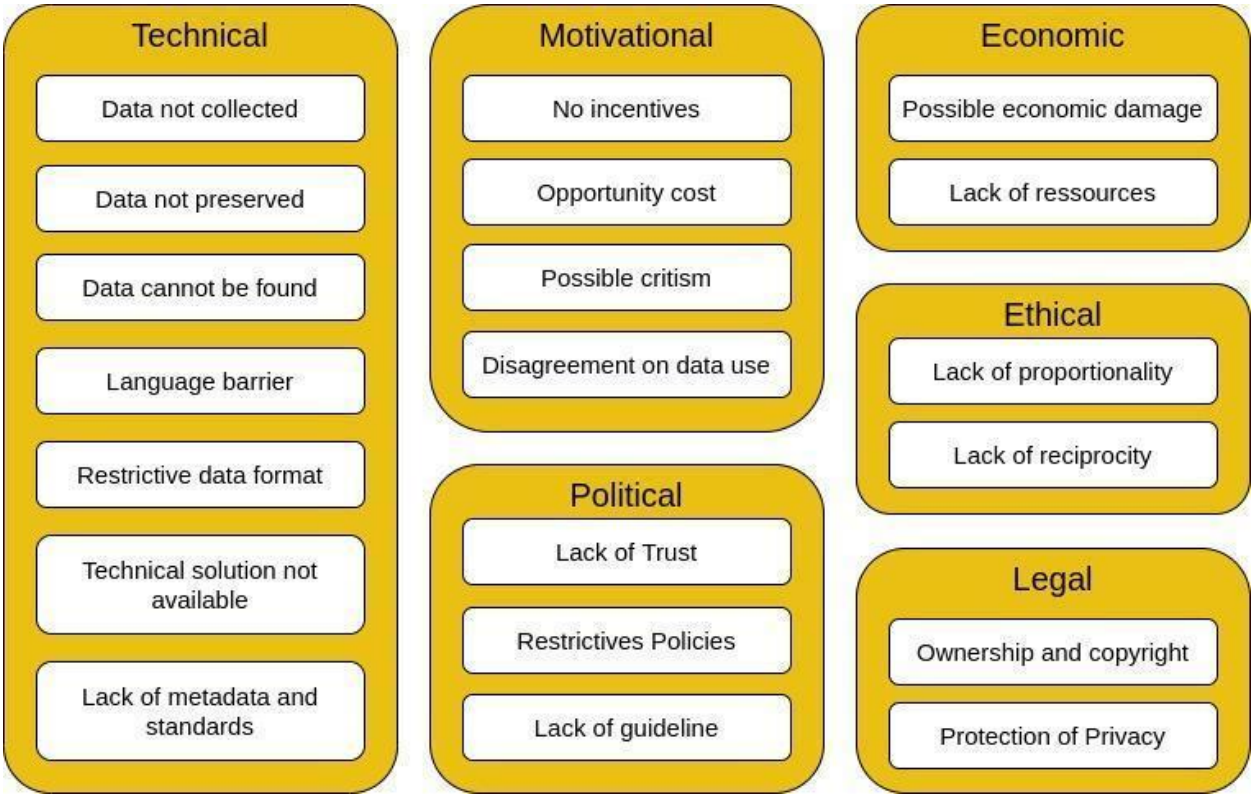


Figure 1: Main barriers for a practical implementation of data sharing taken from *A systematic review of barriers to data sharing in public health*.

Anonymisation process

Through ingestion of reference materials, the workshop and the interviews conducted, we lay out the process and considerations for each step in the methodology.

There is a incremental process within which requires to following steps;

1. Data quality and environment.
2. Data governance and access.
3. Risk analysis.
4. Anonymisation method and tool selection.
5. Quality control.

Some basic high level principles to guide the anonymisation methodology as advised in the introduction of the Anonymizing Health Data book [Anonymizing Health Data]

- The risk of re-identification can be quantified.
- The goldilocks principle: balancing privacy with data utility.
- The re-identification risk needs to be evaluated and then minimised appropriately.
- De-identification involves a mix of technical, contractual and other measures.

Considerations within the anonymisation process

The following is a list of recommended considerations for a de-identification tool taken from both the CFI survey, the workshop and the interviews. It is an iterative process with inputs such as quality control happening along the whole process at frequent touch points.

Prior to tool use

- Consent - Clearly define and establish the consent
- Data Quality (DQ) - DQ to be high and in standardised formatting where possible
- Structure and categorise the text using domain specific Natural Language Processing (NLP) approaches such as developing a data dictionary specific to the data sets [MEDRA Hierarchy]
    - Separate identifiers (direct and indirect if applicable) and the main unstructured text prior to the data moving to the de-identification tool.

    - Use a methodology to structure the text if possible assisted with medical ontologies such as SNOMED CT (a structured clinical vocabulary for use in electronic health records)
- Context - Establish the context and purpose of how the data will be used, the data environment that it will be used in, and factor this in when considering the level of anonymisation.
- Frameworks utilised -
    - 5 Safes Framework. The five safes are managerial and statistical control set of principles and are applied to provide a safe access to the data for research.
    - ADF - Anonymisation decision making framework is a practical guide to anonymisation that gives more operational advice than the ICO's Anonymisation Code of Practice to address a need for a practical guide to GDPR-compliant anonymisation.
- Risk and Privacy -
    - Having a methodology to evaluate the re-identification risk considering the population size, see the statistical framework as explained in the Anonymizing Health Data book (requires purchasing) or any alternative risk assessment (could be provided by external company as part of their tool or system)
    - Define acceptable risk thresholds.

During de-identification

- Quality - The better the quality of the data the increased performance of the tooling
- Ease of use - Tools need to be easy to use otherwise if the process is convoluted it will not be used, or will require training.
- Open-source would allow community contribution and improvement
- Avoid lock-in to a specific platform or programming language.
- Role based access control to the tool and the data.
- Minimised access based on need i.e. reduce joint access to both the tool and the data.
- Meets all regulations relevant for the UK including: Data Protection Act, General Data Protection Regulation, etc. (DPA18 / S171 / GDPR / CLDC / ISO27001 / ISO27034)
- Audit Trail and Associated Protocol for any changes to the data that has occurred
- Validation standard - This is important to ensure the quality of the process (could be provided as part of external tools/automated QC verification systems)

- Continual monitoring - If there is a failure: then a robust feedback loop to identify the gap and to understand the issues
- Adaptability - able to change the tool output in line with the context or level of privacy required

Post tool use

- Privacy assessment once the data has been de-identified.
- Manual Quality Check (QC) and checking along the whole pipeline to maintain frequent touch points.

Anonymisation Decision making Framework

The Anonymisation Decision making Framework (ADF) is an example of a system for developing anonymisation policy and a practical tool for understanding your data situation.

The principles behind the ADF are;

- **Comprehensiveness Principle:** You cannot decide whether or not data are safe to share/release by looking at the data alone, but you still need to look at the data
- **Utility Principle:** Anonymisation is a process to produce safe data but it only makes sense if what you are producing is safe useful data
- **Realistic Risk Principle:** Zero risk is not a realistic possibility if you are to produce useful data
- **Proportionality Principle:** The measures you put in place to manage risk should be proportional to that risk and its likely impact

The ADF has 10 components, which are grouped together into three activities and are explained in the following link *The Anonymisation Decision-Making Framework 2nd Edition 2020*

- A data situation audit (components 1-6). This helps to identify and frame those issues relevant to your data situation. You will encapsulate and systematically describe the data, what you are trying to do with them and the issues thereby raised. A well-conducted data situation audit is the basis for the next core activity.

1. Describe/capture the presenting problem
2. Sketch the data flow
3. Map the properties of the data environment(s)
4. Describe and map the data
5. Engage with stakeholders
6. Evaluate the data situation

- Risk analysis and control (component 7). Here you assemble the processes that you will need to employ in order to both assess and manage the disclosure risk associated with your data situation.

    7. Select and implement the processes you will use to assess and control disclosure risk

- Impact management (components 8-10). Here you consider the measures that should be in place before you share or release data to help you to communicate with key stakeholders, ensure that the risk associated with your data remains negligible going forward, and work out what you should do in the event of an unintended disclosure or security breach.

    8. Maintain stakeholders' trust
    9. Plan what to do if things go wrong
    10. Monitor the data situation

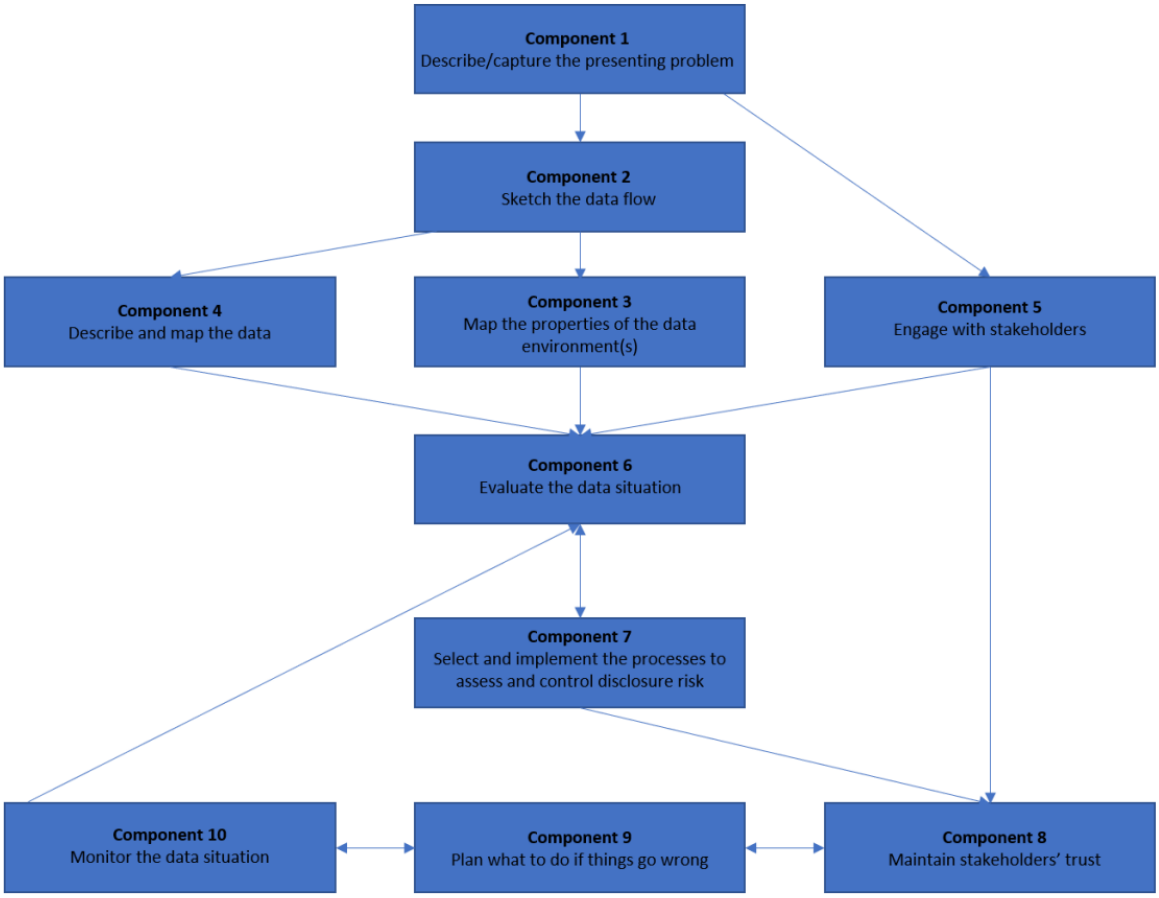


Figure 1: ADF workflow

Assessing the re-identification risk

Personally identifiable information (PII) is one of the most central concepts in information privacy regulation. The scope of privacy laws typically depends on whether PII is involved. The basic assumption behind the applicable laws is that if PII is not involved, then there can be no privacy harm. At the same time, there is no uniform definition of PII in information privacy law.

It has been shown that in many circumstances non-PII can be linked to individuals, and that de-identified data can be re-identified. PII and non-PII are thus not immutable categories, and there is a risk that information deemed non-PII at one time can be transformed into PII at a later juncture. Further problems with non-PII is the increasing availability of information about people. This aspect of the re-identification problem stems from a privacy problem that we will call “aggregation”. Aggregation involves the combination of various pieces of data which can lead to the transformation described.*The PII Problem*

Thus, when we are considering the re-identification risk we need to look beyond PII redaction alone.

Protecting the data

There are three different concepts that envelope data protection and are to be considered when undertaking the protection of the data.

- **Privacy** is about people and personal control about your own spaces and access to your own control and personal autonomy.
- **Confidentiality** is about keeping data in a particular state with assurances of how it will be shared. With healthcare data all NHS staff are bound by the NHS Confidentiality Code of Practice, which sets very clear guidance on how confidential information should be recorded, kept secure, and shared. For guidance on sharing health data the eight Caldicott principles apply to the use of confidential information within health and social care organisations and when such information is shared with other organisations and between individuals, both for individual care and for other purposes.
- **Anonymisation** is the process for keeping those confidentiality assurances when the data is to be used for analytics.

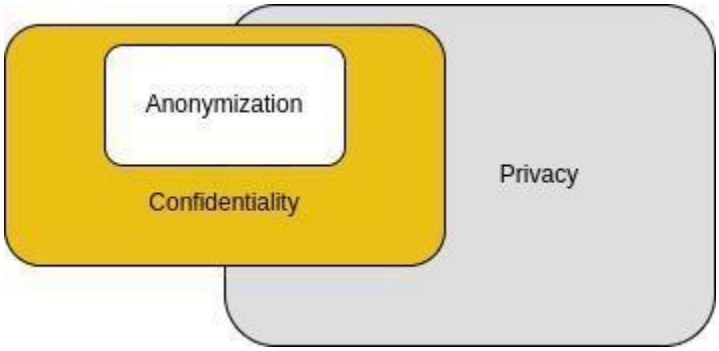


Figure 3 shows the 3 different concepts relating to protecting data.

Methods of re-identification

If data anonymization is not executed correctly or it has weak algorithms within its methodology then re-identification can occur due to the following methods:[blog]

- **Identity disclosure**: Also known as singling out, identity disclosure is the term used to describe situations in which it is possible to identify all or some of the individuals within a dataset.
- **Attribute disclosure**: Attribute disclosure is the ability to determine if an attribute within a dataset is held by a specific individual. For example, an anonymized set of data could show that all employees within the sales department of a particular office arrive after 10 a.m. If it is known that a particular employee is within the sales department of this office, you know that they arrive after 10 a.m., even if their specific identity is masked within the datasets.
- **Linkability**: Linkability refers to when it is possible to connect multiple data points, whether in the same dataset or separate datasets, to create a more cohesive picture of a specific individual.
- **Inference disclosure**: Inference disclosure occurs when you are able to confidently make an inference about the value of an attribute based on other attributes.

Categories of Re-identification attempts

The main threats are stipulated below are taken from Anonymizing Health Data, with the fourth being relevant to public data sets. Each of these categories depends on the context of data release and the controls the sponsor has in place

1. **Deliberate** - Data recipient deliberately attempts to re-identify the data.
    a. Data users using procedural workarounds to de-identified data for benign purposes, perhaps accessing the data through uncontrolled means, e.g. clinical staff pulling PII data from systems to fulfil requirements under time pressures.
    b. Breach of identifiable dataset through user controls meaning that all distributed anonymised data sets become identifiable.
    c. A targeted attack which identifies a high profile individual undermines trust in the system.

2. **Inadvertent** - The data recipient inadvertently (or spontaneously) re-identifies the data.
    a. In controlled environments, ineffective anonymisation of data is common because relevant procedures are not adhered to. The data is then vulnerable to inadvertent reidentification.
    b. Under-representation of patient groups - Anonymisation is less effective for some patient groups which promotes a bias in identifiability risk in such cases.

3. **Data breach** - a data breach at the data recipient's site and the data is "in the wild".
    a. Different versions of the truth begin when a subset of the data is sent in emails and a recipient organisation inappropriately retains the data, often without effectively tracking the changes made.
    b. Lack of common standards for access and approvals encourage distrust across organisations.
    c. Lack of common Extract, Transform and Load (ETL) standards used means that it is hard to assess the level of anonymisation and success of data de-identification, adding to general distrust and reticence to share data.
    d. Ineffective anonymisation in a controlled environment due to poor judgement and unenforced/unknown anonymisation procedures.
    e. Rogue researchers not following protocol and due procedures.

4. **Attack** - An adversary can launch a demonstration attack on the data.
    a. The adversary wants to make a point of showing that a data set can be re-identified
    b. The re-identification is not targeting a specific person, but one or more that are easiest to re-identify
    c. Classified as the worst kind of attack due to producing the highest probability of re-identification
    d. Important features:
        i. Only a single record needs to be re-identified to make the point
        ii. The availability of resources to perform the attack are usually scarce (i.e. limited money)
        iii. Illegal or suspect behaviours will not likely be performed as part of the attack (mostly focus on targeting someone)

Calculating the Risk

Data environment risk analysis - One of the key considerations in this process is understanding the relationship the data has with the data environment it is to be used in and also how it can change over time as this has an impact on the linkage that can be achieved. As access to new data emerges it may affect the data environment with the risk of data linkages and re-identification on previously anonymised data. Therefore frequent assessment of the environment is a step required in the post-anonymisation functions.

Data environments are the set of formal and informal structures, processes, mechanism and agents that either:

a. Act on data;
b. Provide interpretable context for those data or
c. Define, control and/or interact with those data.

ADF Elliot and Mackey (2016)

Calculating the risk of re-identification of patient-level data using a quantitative approach is strongly advised as a preliminary step to take when undertaking the de-identification process. It is the most comprehensive way to show due diligence in the process of data de-identification

and it ensures that the risk is managed and minimised to an acceptable level along with showing that the de-identification is not so excessive as to make the data unusable. The probability of re-identification has been factored into statistical analysis and software has been developed by commercial companies so that this calculation can be automated. The whole data set is assigned a probability value either using maximum risk, where an adversary is trying to re-identify a single person or an average risk approach where an adversary is trying to identify an acquaintance.

Overall risk of an attempt - Once all risk elements have been calculated, they are combined to produce a comprehensive set of metrics describing the level of de-identification and risks of de-identification of the entire data set. The first step is to derive the overall risk of re-identification which is the product of risk of attempt and risk of successful re-identification of records in the set. See below for the equation and further guidance can be found at Calculating the Risk of Re-Identification of Patient-Level Data Using Quantitative Approach by PhUSE.

pr(identification)= pr(attempt) x pr(identification/attempt) *marsh et al 1991*

NOTE pr = probability

How can we mitigate these risks/threats?

1. Implement known procedures and enforceable contracts in place with the data recipient.

2. VIP flags used to identify high profile individuals, e.g. Prime Minister's family, where a pseudonym is attached to the record.

3. Access restrictions
    a. Secure environment with effective controls in places.
    b. Minimise multi-layer access i.e. to data and tooling.
    c. Ensuring easy access to those that are authorised to access the data, to increase efficiency.

4. De-identification risk analysis on the data set which considers the data environment. i.e. publicly available data sets and the possibility of linkage to other data sets.

5. Regular updates to tools and risk assessment as data environments can change as greater access to open data becomes available.

6. Maintain situational awareness of current threats and best practice through coordinated focus groups and to facilitate engagement.

7. Impose laws, sanctions and fines as stipulated in the Data Protection Act which is enforceable by the ICO (Information Commissioner's Office). It states that a contravention can result in substantial fines being levied against offending personnel and companies.
    a. £180-200 per user, per breach.
    b. Financial - ICO fine.
    c. Lawful prosecution
    d. Audit tracking.

8. Frequent manual human touch points throughout the process to maintain quality and to ensure no data breaches.
    a. Build trust and appoint Point Of Contact (POC) if required.
    b. Security training.
    c. Encourage positive behaviours.

9. Methodology to report faults or near misses.

10. Audit trail of the whole process and any changes time stamped along with documentation of the responsible person (if it is a manual step). Automation of this is suggested with manual overarching QC.

Anonymisation levels

The aim of de-identification is to reduce the risk of re-identification to an acceptable level while retaining as much data utility as possible (functional). This is in contrast to removing the risk altogether which would inevitably deem the data unusable (absolute). Calculating the risk of re-identification should consider the likelihood of making a distinguishable line between 'could' and 'will' and the impact that it will have so that the risk is managed but not eradicated.

The 4 levels of anonymisation.

1. **Absolute anonymisation** - zero possibility of reidentification under any circumstances (effectively having data which is of no use)
2. **Formal anonymisation** - de-identification i.e. stripping away of the direct identifiers (including pseudonymisation)
3. **Statistical anonymisation** - attempts to measure the risk of a re-identification happening and to control the risk. Focussed heavily on the properties of the data.
4. **Functional anonymisation** - acknowledges the value of the statistical approach but focuses on the value of the data, acknowledges the statistical significance but also takes into account the environment in which the data exists.

Establishing the requirement of anonymisation level

The levels of anonymisation of data must match the utility required by the end-user and hence, it is imperative to establish the context in which the data is to be used, and the external access controls to which will be applied, in order to ascertain the level of anonymisation required.

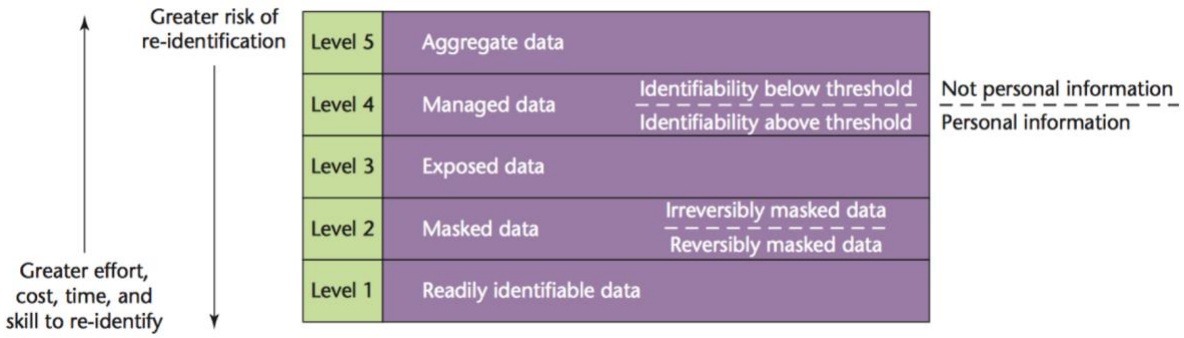


Figure 4. Showing the goal of de-identification and the risks associated with each level taken from *"Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification*

Explanation of the five levels of disclosure control.

- ***Level 5**: Aggregate data that cannot identify individuals. Through the use of aggregation methods, non-stratified counts, frequencies or rates are shared. Note that not all aggregated data meets this requirement if the cell size (sample size of the group being selected by the combination of filters) for a given crossing of some combination of variables can lead someone to identify a particular individual. An example might be people who responded to a survey where the sample size of one particular variable – say race – is small enough to deduce who that individual might be (with or without additional data.)*

- ***Level 4:** Managed data. This is where the researcher actively manages (and measures) the degree to which re-identification can occur. If the risk is low (according to an established benchmark) then the data is considered managed with or without personal information being considered identifiable.*

- ***Level 3:** Masked identifiers and non-identifiers. As with Level 2, the identifiers (such as name and date of birth) are masked, but with Level 3, we also mask variables that are considered to be quasi-identifiers. An example might be that we mask gender, but fail to mask the variable that contains the last date of a pap smear or pregnancy flag.*

- ***Level 2:*** *Data that is masked or obscured. For example, you may modify the "identifying" variables through randomization and creating reversible or irreversible pseudonyms. Most BioPharma companies conducting clinical trials use this technique to mask treatment or intervention information.*

- ***Level 1:*** *Data that's clearly identifiable. For example, a database containing sensitive PII names such as Social Security Numbers (for the US but any unique identifier for example the 18 requirements in HIPAA) , or non-sensitive PII such as dates of birth or other identifying information that can be easily accessed publicly.*

Privacy type and data utility

There are different types of privacy methodologies. A simple de-identification technique is to retain aggregate data, and delete all underlying records, however, this can only be achieved for data which is already structured. As a result, no data is individually identifiable as it is made up of many rows and represents only totals, such as counts, averages or correlations. The following table summarises nine different types of privacy techniques which can be applied to **unstructured data** and shows the privacy level and data utility associated with it. These privacy types can get combined together to reach a specific outcome such as using both Differential Privacy and Federated Learning for example.

Type of Privacy	Adoption Stage	Ease of use	Speed and Size	Privacy Level	Data Utility	Underlying data
Process and Administrative Control	**Widely adopted**	**Medium**	**Medium**	**Medium**	**High**	**Remains**
	Typical Concern		**Raw data is still present and could be misused by malicious actor**			
Aggregation	**Widely adopted**	**Simple**	**Fast/large**	**Highest**	**Low**	**Removed**
	Typical Concern		**Limited additional information can be derived**			
Redaction	**Majority adopting**	**Simple**	**Fast/large**	**High**	**Medium**	**Partially removed**

	Typical Concern		Redaction may be difficult to apply: remove too much utility			
Hashing and Linking (or, "Pseudonymisation")	Majority adopting	Simple	Fast/large	Highest	High	Removed
	Typical Concern		Does a mapping back to the identifiable information exist ?			
Synthetic Data	Early adopters	Medium	Medium	Highest	Medium	Removed
	Typical Concern		Is the resulting data accurate across all dimensional queries			
Differential Privacy	Early adopters	Medium	Fast/large	High	Medium	Partially removed
	Typical Concern		Is the resulting data accurate across all dimensional queries			
Federated Learning	Innovator	Complex	Medium	Highest	Low	Not shared
	Typical Concern		Can the model be obtained without data being linked or sent			
Homomorphic encryption	Innovator	Complex	Slow/Small	Highest	High	Encrypted
	Typical Concern		Computing prescription adherence across multiple pharmacies			
Multiparty computing	Innovator	Complex	Slow/Small	Highest	High	Bot shared
	Typical Concern		Is the data set too big or the query too complex ? Will exploratory analysis be required ?			

Anonymisation tools

The process of de-identifying a dataset requires the application of different techniques. The main ones, used extensively in healthcare, that have been acceptable for data analysts are;

1. Generalisation - Reducing the precision of a field
2. Suppression - Replacing a value in a data set with a NULL value
3. Subsampling - Releasing a simple random sample of the data set.

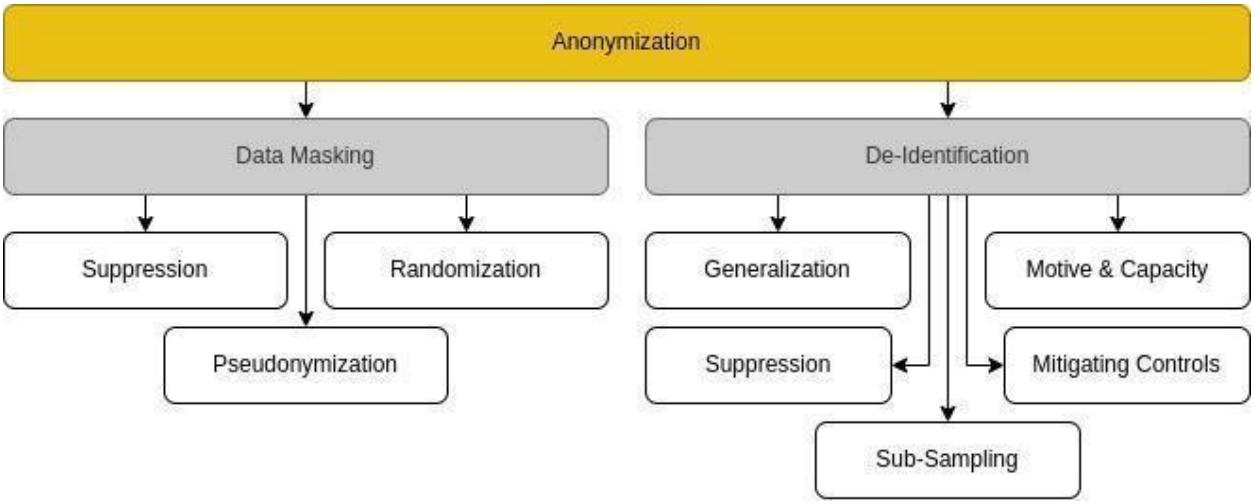


Figure 5. Anonymisation methods classified into two classes, data masking and de-identification Ref

Generalisation There are many different types of generalisation techniques. Generalisation is the process by which personal identifiers, including names and other information that someone could use to identify a person are removed from data while still preserving its relevance for an analysis or research purpose.

Suppression. Replacing a value in a data set with a NULL value (or whatever the data set uses to indicate a missing value).

Subsampling. Releasing only a sample of the data set rather than the whole data set.

Tools in use

The following table lays out a non-exhaustive list of the tools in use as captured in February 2022. These are all healthcare related and the URL is either for the website associated with the tool or to the paper explaining the approach for the tool. The latter part of the table are those that were mentioned at the workshop. The tools listed tend to focus on one context e.g. patient data, doctor notes, etc. and therefore you would have to change the tool depending on text input.

Name	Type	Free text	Risk Analysis	Compliance	Comments
Scrub System	Proprietary	X			Datafly became Scrub
Datafly System	Proprietary				
µ-Argus System	Proprietary		X		
Facit data Systems	Proprietary	X		GDPR	
Privatar	Proprietary	X		HIPAA / GDPR	
Privacy Analytics	Proprietary	X	X	HIPAA / GDPR	
Phemi	Proprietary				Full solution Not only de-id
Imperva	Proprietary			HIPAA / GDPR	
InCountry	Proprietary			HIPAA	
Immuta	Proprietary	X			
BigID	Proprietary			GDPR	Full solution Not only de-id
Mirador Analytics	Proprietary	X	X	HIPAA	
OneTrust	Proprietary	X			
Presidio	Open-source	X			
deid	Open-source	X		HIPAA	
Philter	Open-source	X			
Mask	Open-source	X		HIPAA / GDPR	
From Workshop					

Cognition de-identification	Open-source	X			
Cogstack - SemEHR	Open-source	X			
Cogstack - MedCAT	Open-source	X			Only medical terms
Spacy EntityRecognizer	Open-source	X			Separate toolbox per type of text
UMLS Ontology	Licence				Free-text to structured data
AWS Comprehend Medical	Proprietary	X			
QualiAnon	Addon	X			
CRIS The Clinical Record Interactive Search		X		GDPR	Developed by SLAM Biomedical Research Centre (BRC) to enable routinely collected EHRs to be used in research, using an explicit de-identification process

Challenges to anonymisation

The following is an output of the project workshop held on 2nd March 2022 which gathered together experienced personnel in this field, identified through research papers, departments leads and commercial providers. The aim was to discuss the issue of maintaining privacy of individuals identified in unstructured health data whilst maximising its utility for analysis and research. Challenges to anonymisation are also discussed in the following link which is useful to corroborate the workshop output. "A systematic review of barriers to data sharing in public health"

1. **Lack of collaboration/research**
   a. Evidence and research to prove the impact of data redaction on research outcomes.
   b. Need to evidence anonymisation process balance with local providers e.g. ICS and Trust level operate differently.
   c. Governance/Access requirements being too strict resulting in apathy and disengagement.
   d. User greed or excessive demands (e.g. they don't want anonymisation).
   e. Organisations with developed domain-specific dictionaries used for the tooling resisting sharing and collaboration.

2. **No guidelines**
   a. Lack of common de-identification standards and definitions
   b. Lack of guidelines or methodology process if people do not have the deep knowledge
   c. Burden of manual anonymisation
   d. Lack of resources for the testing and quality assurance of the automated outputs
   e. Lack of benchmarks for 'good enough' (basic level to attain to)
   f. Domain specific: lack of comprehensive privacy dictionaries, ontologies, etc... not only context-aware.

3. **Excessive risk aversion.**
   a. Lack of sufficient knowledge and guidelines meaning people are not confident in their protection against data breaches.

b. Setting the risk using a general risk approach rather than considering the specifics of the data and possible attacks

4. Lack of IT capability and skilled personnel to undertake the de-identification

a. Tool development, de-identification and ongoing Quality Assurance (QA) of methods.
b. User preferences: there needs to be different levels for experts or laymen.
c. Time and resource limitations in having to adapt and accommodate to different anonymization standards.

5. Data

a. **Structuring via s**tandard terminologies can constrict clinician's intended meaning
b. Population size is too small e.g. rare diseases/specialist care such as a cluster of children tested at Great Ormond Street Hospital (GOSH) for specialist bone tumours were identifiable (in part due to social media presence) within Local Authority (LA) data
c. Level of sensitivity of the data classification: the degree of sensitivity of a cold/flu might be substantially different from that of cancer as perceived by the individual with that condition
d. Language issues: simple text analysis often relies on things like proper sentence structures
e. Big Data availability: with the rise of greater amounts of publicly available information, especially open data, the linking of seemingly harmless information with side information could result in privacy breaches of PRI
f. Buildup of information: some information may not be privacy revealing if only a small quantity of data is considered, but may be sensitive in larger amounts

6. User concerns

a. Patients' reluctance for data to be shared (anonymised or otherwise)
b. Lack of clarity about processes > end user workarounds
c. Concerns around sharing within and outside of NHS
d. Default closed attitude

7. Data quality

a. Data Quality - analysts sometimes don't trust Data Quality (DQ) in anonymised datasets due to lack of transparency and want access to raw data
b. Also can question data linkage approaches and quality again often due to lack of transparency (done by someone else pre-anonymisation)

8. Poor tool performance

a. The large free-text datasets used for training models come from internet sources such as Wikipedia, blogging sites and social media, so are not healthcare focussed in terms of vocabulary and context. There is also a US-English bias in many pre-trained Natural Language Processing (NLP) language models
b. Concerns about bias - need demographic details to ensure fair data.
c. Performance on non anglo-saxon names

9. Big picture effects

a. Utility of data in anonymised state is affected
b. Linkage across services is lost

Case studies and examples

An example of a current de-identification practice

The following de-identification process was presented during the workshop.

Roles involved

The process had access to a 'security team' of advisors which contained the following roles;

- Patient leads
- Caldicott advisors
- IG solution experts
- SLAM reps
- Child protection
- Clinical representatives
- Academic representatives

Method

The approach is based on building a patient-level dictionary containing all patient identifiers such as name, date of birth, NHS Number, medical terminology but these were able to be built from linked structured data available alongside the unstructured data. This dictionary is then used to de-identify free text in the patient record from any of the identifiers present in the dictionary. Concerning postcodes, these are truncated to the Lower Layer Super Output Area (LSOA) area and birthdays are rounded to the first of the month. This approach has the advantage of being reliable and effective as the data needed to be masked is already identified. In order to match a specific pattern, regular expressions (regex) are used to search the unstructured data and discovered identifiers are replaced by "XXXXXXX" or a pseudonym. This allows as much as possible of the clinical part of the text to be kept intact for research purposes.

However this approach has its limitations on de-identification, such as any other names present in the clinical free text or misspelt words that are not in the patient level dictionary, can be overlooked and not masked. Thus to further mitigate risk on the shared data, it is then required to get someone with authorised access to review the data. This requires an assessment as to the scale of the data and the appropriate methodology to use, whether it is a 'dip-test' (review

of a random sample) or an assessment of every document (as completed by the company Privacy Analytics in their methodology). Also for this example, an audit log of all data searches is assessed weekly by a board to ensure that the data accessed is in line with the research in progress.

Tool used

--- REDACTED ---

The pros of this methodology was that it has good performance and high precision but the cons were that it required a high metadata input; it was patient specific and had the possibility of a 'nosy neighbour' breach.

Shared understanding

The operational side is overseen by a patient led committee built on 14 years of communication to establish trust within the community. For transparency, the explanation of 'why' given research is being conducted is always offered.

Examples of privatisation failures

1. Manual privatisation

a. Social Security Number leak: In 2008, Public.Resource.org released an audit of Public Access to Court Electronic Records (PACER) documents and noted 1,600 cases in which litigants submitted documents with unredacted Social Security numbers, and many actions where the redaction was performed incorrectly by simply placing a black box on top of the taxpayer ID, leaving the numbers untouched underneath the graphic. It was as simple as the copy/paste technique (ctrl+A, ctrl+C, ctrl+V) to reveal the “redacted” text.

b. Redaction table listed at the end of the document in alphabetical order. Adversaries can find the starting letter and deduce the name of people via cross-referencing with News/Social Media/…

c. Ghislaine Maxwell Deposition example - the deposition includes a complete alphabetised index of the redacted and un-redacted words that appear in the document, which was able to be re-associated with their locations in the redacted text.

2. Small population size re-identification

a. Cross matching references when the data is “too specific”. For example, individuals of incredibly rare diseases with small patient populations. In the case of a new strain of disease or new medical solution being used to treat a patient, any sharing of data, whilst necessary for the development of further medical provision across the NHS, is impossible to do without in-turn identifying the individual’s personal circumstances.

b. Spontaneous recognition - you know of person X who has an unusual combination of attribute values. You are working on a data set and observe that a record within that data set also has those same attribute values. You infer that the record must be that of person X. In order to be truly spontaneous you must have no intent to identify *Statistical Confidentiality: Principles and Practice*. Disclosure from micro data sets is often possible, and difficult to prevent, unless the information in the data set is severely reduced. Disclosure of "rare persons" can be prevented by taking care of the unique in two- or three-dimensional tables. The

probability of disclosure by 'response knowledge' can be limited by advising the respondents not to tell anyone else that they were in the survey.

For example, if we take a data file with four variables: household composition (H) in 24 categories, age (A) in 14 categories, marital status (M) in 2 categories, and sex (S) in 2 categories. The critical population sizes were obtained using the relative criterion value of 0.1%; that is, the number of possibly identifiable records in any subpopulation must be smaller than 0.1%. In a population of 14,000,000 (the Dutch population) this means that 14,000 people were unique. If someone is unique in the population, the following question may arise: How high is the risk of identification? This depends on the amount of knowledge available to some users of the data. It is difficult, but not impossible, to model additional knowledge and to quantify the probability that someone has knowledge of certain information. We think disclosure protection for this kind of malpractice could and should be taken care of by legal arrangements, and not by restrictions on the data to be released. *Disclosure Control of Microdata*

3. Machine Learning

a. Generative Pre-trained Transformer 2 (GPT-2) - issues may arise if a generative model trained on private data were to be made publicly available because they can sometimes reproduce sensitive data, including personally identifiable information (PII) such as names, phone numbers, addresses, etc. Training datasets can be large (hundreds of gigabytes) and pull from a range of sources, be that private or public data. This raises the possibility that a model trained using such data could reflect some of these private details in its output. It is therefore important to identify and minimise the risks of such leaks, and to develop strategies to address the issue for future models. Privacy considerations

Re-identification examples

Anonymisation standards - *Estimating the success of re-identifications in incomplete datasets using generative models* This report suggests that even heavily sampled anonymised datasets are unlikely to satisfy the modern standards for anonymisation set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model. When a data administrator practises release-and-forget anonymisation, this means that records are released – either publicly, privately to a third party, or internally within their own organisation – and then forgotten, meaning there is no tracking on what happens to the records after release. Broken Promises of Privacy: Responding to the surprising failure of anonymization

Re-identification attack success rate - In *A Systematic Review* the overall success rate for all re-identification attacks was estimated as between 26% and 34% for health data. However, these results mask a more nuanced picture that makes it difficult to draw strong conclusions about the ease of re-identification. The confidence interval around the above estimates was large, partially because many of the attacks were on small databases. Therefore, there is considerable uncertainty around these numbers. We found only two studies where the original data was de-identified using current standards and for those the data was successfully re-identified. Only one of these attacks was on health data, and the percentage of records re-identified was 0.013%, which would be considered a very low success rate.

High cost of motivated attack - Evidence to suggest the time and resources involved in a motivated attack is often costly and hence the probability of this type of attack is low. This report highlights the effort that was required to re-identify a data set "*Six suspected matches with low confidence scores were identified. Each suspected match took 24.2hrs of effort. Social media and death records provided the most useful information for getting the suspected matches*"

How non-PII can be linked to re-identify PII

Computer scientists are finding ever more inventive ways to combine various pieces of non-PII to make them PII.

Non-PII data? - According to a study done by computer science professor Latanya Sweeney, the combination of a ZIP code, birth date, and gender will be sufficient to identify 87% of

individuals in the United States. These pieces of data are all generally considered to be non-sensitive PII but when combined can still cause a risk of re-identification.

Localised risk assessment is a requirement. The percentage of a US state's population estimated to be vulnerable to unique re-identification when protected via Safe Harbour and Limited Datasets ranges from 0.01% to 0.25% and 10% to 60%, respectively. Evaluating re-identification risks This illustrates that blanket protection policies, such as Safe Harbour, leave different organisations vulnerable to re-identification at different rates. It provides justification for locally performed re-identification risk estimates prior to sharing data.

Netflix Reviews - Narayanan and Shmatikov found a way to link this data with the movie ratings that participating individuals gave to films in the Internet Movie Database (IMDb), a popular website with information and ratings about movies. They concluded: "Given a user's public IMDb ratings, which the user posted voluntarily to selectively reveal some of his . . . movie likes and dislikes, we discover all the ratings that he entered privately into the Netflix system, presumably expecting that they will remain private." One user had posted an identifiable review on a different website and this privacy revealing information revealed their orientation - they successfully sued Netflix. *De-anonymization of Netflix Reviews using Amazon Reviews* report the conclusion is that removing personal identifiable information from datasets is not enough to anonymise data. Through the use of exact matches and similarity scores, we were able to identify Netflix users in the dataset (by first and last name), using only publicly available Amazon reviews. A main takeaway of this project is that companies need to be careful when releasing any user data, and must take into account all data available on the internet in addition to the data being released. When companies release compromising information, it can harm their trustworthiness, which could have serious implications for their business.

AOL - In 2006, AOL released twenty million search queries for the benefit of researchers. These queries were considered to be fully anonymised. Yet, reporters from the New York Times quickly demonstrated that at least some of this information could easily be re-identified. The reporters showed how they were able to identify one person based on her search queries—User No. 4417749.

Example of a healthcare paper focussing on anonymisation

The following example has been extracted from the paper: Experimentation with Personal Identifiable Information - Sabah Al-Fedaghi, Abdul Aziz Rashid Al-Azmi. It identifies and classifies PII into different categories such as Atomic PII (APII) is information that has a single human referent, APII is said to be self-APII (SPII) if its subject is its proprietor and only its proprietor. Singleton SPII (SSPII) if it is an SPII such that its proprietor is the only entity involved; otherwise, SPII is called a multitude SPII (MSPII). CPII: Compound PII (CPII) is information that has more than one human referent. SSPII, MSPII and CPII are presented in the diagram in Figure 3 of the paper.

> Psychiatric Report
>
> Since receiving the diagnosis of neural tumor (1), "Lucy" has felt depressed and anxious about her health (2). Lucy has experienced two nights of restless sleep (3). She has lost enthusiasm for her usual activities, such as going shopping (4) and taking care of her son "Tim" (5). She reports having no energy for maintaining her work or social life (6). She has also become more irritable and aggressive (7), which is putting additional pressure on her family (8). She admits to being preoccupied with thinking about her illness (9) and is having trouble concentrating on daily activities (10). She reports feeling tired (11) but too scared to sleep for fear that she will not wake in the morning (12). In conjunction with her depressive symptoms, Lucy is also experiencing excessive anxiety (13). Her anxiety is associated with restlessness, tiredness, irritability, insomnia, and difficulty in concentrating (14). Other symptoms include palpitations, tachycardia and flushing (15). Lucy expresses concern over the impending biopsy report (16), due sometime in the next two days, asking "Am I going to die? Does the tumor mean cancer?" (17). Lucy also expresses concern over her son's welfare while she is hospitalized (18). In the last month, her fiancé James and her mother Hermione have been looking after both her and her son (19).

The diagram shown after the table on the following page is used as a visual aid for privacy officers to identify and decide on the sensitivity of the PII. For example, "Hospitalised", "Aggressive" and "Depressed" might be the most sensitive compared to "Tired", or "Anxious". Table 1 from this paper is attributing a sensitivity coefficient to the verb and the rest-of-PII to determine a combined sensitivity equation. This shows how context can affect the decision about which words are sensitive.

Table 1. Levels of sensitivity of the psychiatric report.

PII #	Verbs	Verbs Sensitivity	Rest-of-PII	Rest Sensitivity	Combined Sensitivity Equation (1)	Combined Sensitivity Equation (2)
1	receive	3	diagnosis of a tumor	9	12	27
2	feel	5	depression, anxiety	5	10	25
3	experience	1	restless sleep	10	11	10
4	lose	6	enthusiasm	4	10	24
5	take care	1	caring for her son	2	3	2
6	having	1	maintain her life	5	6	5
7	becomes	3	more aggressive	8	11	24
8	put	1	pressure on family	10	11	10
9	admit	6	about illness	6	12	36
10	having, concentrating	4	effect on daily activities	2	6	8
11	report	2	tired	8	10	16
12	scared, sleep	6	dying during sleep	5	11	30
13	experience	2	excessive anxiety	9	11	18
14	associate	1	symptoms on her behavior	10	11	10
15	include	2	physical symptoms	10	12	20
16	express	5	impeding report	1	6	5
17	ask	3	tumor and cancer	10	11	30
18	express	7	son's welfare	4	11	28
19	look after	2	family members	7	9	14

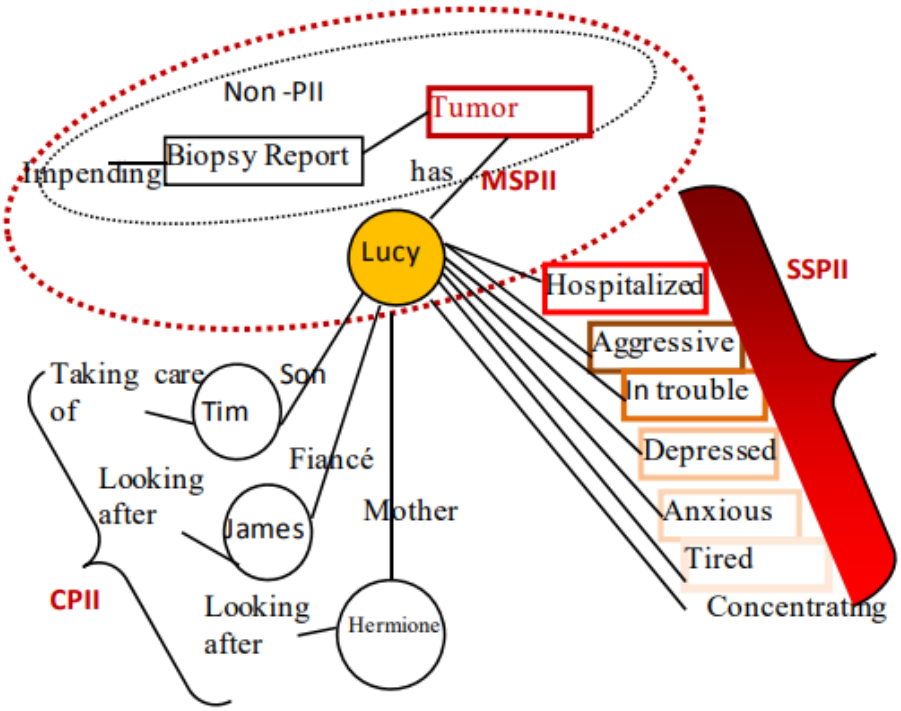


Figure 4. Lucy's PII sphere.

List of key qualities for a tool

1. **Structuring and data handling**
    a. Ability to flag and identify with the range of possible data issues prior to de-identification (misspellings, medical terms, acronyms)
    b. Connection with a clinical vocabulary in order to match and assist word identification to assist structuring of the data.
    c. Ability to flag the data variables required for anonymisation to assist in the risk analysis and disclosure control process
    d. Ability to deal with unstructured, semi-structured and structured data
    e. Ability to deal with different formats of free-text data e.g. medical notes, patient feedback, survey responses, research papers

2. **Tool Use & Validation**
    a. Ease of manual manipulation in order to react to the level of anonymisation required and the key variables to be maintained for data utility
    b. Automated auditing of the flagged terms, any data manipulation and tool manipulation that has taken place
    c. Ability to demonstrate quality and anonymisation level before and after each stage of the de-identification process for the QC process.
    d. Ability to apply manual QC at each step along the process QC (human and automated) or the requirement of human authorisation to move to the next step
    e. Clarity around the tool limitations
    f. Need to align with the Information Governance (IG) process

3. **Context**
    a. Ability to tune into a domain extracting and utilising the appropriate medical dictionary.
    b. Clarity around individual versus population
    c. Ability to define level of anonymisation

4. **Flexibility**
    a. Ability to adapt the anonymisation functionality to the risk level assessed
    b. Flexibility within the tool programming to adapt to the utility required and hence the purpose of the output data aligning the appropriate level of de-identification
    c. Incorporated regular updating and reaction to current “threats”

Recommendations for future work

Knowledge

1. **Convene** ongoing quarterly round-table of identified stakeholders to review progress towards the UK equivalent of Coordinated Access for Data, Research and Environments (CADRE) or similar.
2. **Contact** specific patient representative stakeholders for survey responses as they were unable to attend or respond within the timescale for this phase of work.
3. **Attend** HealTAC 2022: the 5th Healthcare Text Analytics Conference (15-16th of June 2022) in Edinburgh
4. **Understand** the levels of structured data, what does it mean and how can it be achieved?

Assess

5. **Liaise** with the data users and drill into their specific requirements of the data as potentially they may just require synthetic or structured data which is easier to risk-assess the re-identification risk.

Tooling

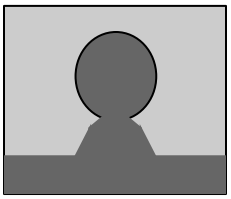
6. **Research** into how the data can be semi-structured as the baseline for risk analysis and analytical use.
7. If funded, **commission** a follow up piece of work. Gather detailed requirements for the tool and assess which of the tools listed in this report best meet the requirements.
8. If none are suitable, **conduct a gap analysis** and approach the developers to ask whether additional functionality could be added to a tool to meet requirements or whether it's better to begin development of a new tool.
9. **Consider synthetic data development** based on real data in order to allow secure analysis or for future training of ML tools.
10. Further **research and utilise tools on the market** to understand the capabilities and the possibilities for own development

Support

11. **Reach out to SME's** on the contact list for further support

12. **Identify roles required**, advisors and patient voice representatives for a whole team approach.
13. **Utilise this report** as a host for discussion, disseminating it to Policy colleagues and the wider group of interested academics to obtain viewpoints and advice.

About The Authors



NHSE Digital Analytics Research Team (DART) (previously NHSX Analytics Unit) supports the analytical work within the Transformation directorate of NHSEI through delivering programme analytics, supporting the professionation of the NHS analytical community and leading on innovation to get more value out of NHS data

Sara Boltman is the founder of Butterfly Data and specialist in data science, intelligence analysis, and management consultancy. Taking unstructured data and extracting meaning and intelligence from it then presenting this to senior decision makers to inform their strategy is her core strength.

Thomas Doublein is a specialist in machine learning handling large volumes of structured and unstructured data. Experience of developing algorithms for classification, training and testing multiple candidate models and cross-validating to ensure appropriate balance between false positives and false negatives. Open source software developer.

Sally Wrigley is a Scrum master and Delivery lead with a background in project management, coaching and instructing. She covers a multitude of disciplines, with prominent expertise in stakeholder management and business analysis. A captain in the Army Reserves, Sally maintains an eye for detail and keen organisational skills.

Butterfly Projects Limited, Trading as 'Butterfly Data'

Company Registration No: 04952566 - VAT Registration No: 850986489

253 Cowbridge Road West, Cardiff, CF5 5TD

info@butterflydata.co.uk | www.butterflydata.co.uk

Annex 1 - Terminology Definitions

It should be noted that terminology can have variance in different pieces of legislation and across different jurisdictions and across countries therefore we have defined the terminologies we are using in this report below.

Personal data means any information relating to an identified or identifiable natural person ('data subject'). An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.[ICO-what is personal data?]

Direct identifiers are fields in a data set that can be used alone to uniquely identify individual or their households e.g. name, unique ID number (NHS, Passport, Driver licence)

Indirect identifiers are fields in the data set that can be used in combination with another to identify individuals. E.g. date of birth and geographic location.

Indirect identification means you cannot identify an individual through the information you are processing alone, but you may be able to by using other information you hold or information you can reasonably access from another source. E.g individual's licence plate number. The police (a third party) can quickly match a name to a licence plate number.

De-identification is the process of removing or obscuring the direct-identifiers in some way preventing somebody being able to recognise an individual directly from the data. [M.Elliot]

Anonymisation is the process of removing or obscuring the direct and indirect identifiers, preventing someone from recognising an individual indirectly from the data and other information. It is a process by which personal data is rendered non-personal defined in an ISO standard (ISO 29100:2011) as the 'process by which personally identifiable information (PII) is irreversibly altered in such a way that a PII subject can no longer be identified directly or indirectly, either by the PII controller alone or in collaboration with any other party [Anonymisation and GDPR compliance]. If any reasonably available means would enable re-identification of the individuals to which the data refers, such data would have been ineffectively anonymised and merely

pseudonymised. This means that despite an attempt to anonymise the data, personal data is still being processed.[ICO-what is personal data?]
Therefore in the risk process we need to decide whether a set of indirect-identifiers are sufficient to be able for somebody to be identified.

De-identification techniques refer to methods used for transforming a dataset with the objective of reducing the extent to which information is able to be associated with individual data subjects [O/IEC 20889:2018(en)]

Personally Identifiable Information (PII) refers to information that permits the identity of an individual to be directly or indirectly inferred, including any information that is linked or linkable to that individual.

Privacy Sensitive Information (PSI) refers to information that, depending on the user's perception, has the consequence of revealing the privacy of the individual.[Challenges in detecting privacy]

Privacy Revealing Information (PRI) refers to the superset of PII and PSI.

Data Re-identification or de-anonymisation is the practice of matching anonymous data (also known as de-identified data) with publicly available information, or auxiliary data, in order to discover the individual to which the data belong

Safe Harbour policy enumerates 18 identifiers that must be removed from health data, including personal names, web addresses, and telephone numbers. This process creates a public use dataset, such that once data has been de-identified under this policy, there are no restrictions on its use. As in many data sharing regulations in the USA and around the world, Safe Harbour contains a special threshold provision for geographic area. When a geographic area (eg, zip code) contains at least 20 000 people, it may be included in Safe Harbour protected datasets, otherwise it must be removed.

Semi-structured data refers to what would normally be considered unstructured data, but that also has metadata that identifies certain characteristics. The metadata contains enough information to enable the data to be more efficiently catalogued, searched, and analysed than strictly unstructured data.

Structured data is data that has been predefined and formatted to a set structure before being placed in data storage, which is often referred to as schema-on-write.

Annex 2 - References

Name	URL	Notes
ICO what is personal data	https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/what-is-personal-data/#pd5	Standard definitions and regulations governing UK data handling
The Five Safes is a framework for helping make decisions about making effective use of data which is confidential or sensitive.	http://www.fivesafes.org/	Developed by Felix Ritchie in 2002/3 at ONS
10 Misunderstandings related to anonymisation	https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf	
ISO/IEC 20889:2018(en) Privacy enhancing data de-identification terminology and classification of techniques	https://www.iso.org/obp/ui/#iso:std:iso-iec:20889:ed-1:v1:en	International standards relating to privacy
Anonymisation and GDPR compliance	https://www.gdprsummary.com/anonymization-and-gdpr/	
Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification - Gregory S. Nelson	https://www.researchgate.net/publication/318866074_Practical_Implications_of_Sharing_Data_A_Primer_on_Data_Privacy_Anonymization_and_De-Identification	
THE PII PROBLEM: PRIVACY AND A NEW CONCEPT OF PERSONALLY IDENTIFIABLE INFORMATION	https://www.law.berkeley.edu/files/bclt_Schwartz-Solove_NYU_Final_Print.pdf	
Challenges in detecting privacy revealing information in unstructured text	*http://ceur-ws.org/Vol-1750/paper-05.pdf*	
Breach Report – DEFRA – UK Statistics on Waste – July 2021	https://uksa.statisticsauthority.gov.uk/publication/breach-report-defra-uk-statistics-on-waste-july-2021/#pid-corrective-actions-taken-or-planned-to-prevent-re-occurrence	Example of a data breach, the impact and how it was dealt with.
Montreal Accord on Patient-Reported	https://pubmed.ncbi.nlm.nih.gov/28	

Outcomes (PROs) use series - Paper 9: anonymization and ethics considerations for capturing and sharing patient reported outcomes	433677/	
MedDRA Hierarchy (medical dictionary for regulatory activities)	https://www.meddra.org/how-to-use/basics/hierarchy	Medical dictionary for regulatory activities International a rich and highly specific standardised medical terminology to facilitate sharing of regulatory information internationally for medical products used by humans
Data protection Act 2018 sect 171	https://www.legislation.gov.uk/ukpga/2018/12/section/171?view=plain	It is an offence for a person knowingly or recklessly to re-identify information that is de-identified personal data without the consent of the controller responsible for de-identifying the personal data.
An overview of the Data protection act 2018	https://ico.org.uk/media/for-organisations/documents/2614158/ico-introduction-to-the-data-protection-bill.pdf	Stipulates the fines and the levels of enforcement
Calculating the Risk of Re-Identification of Patient-Level Data Using Quantitative Approach	https://www.lexjansen.com/phuse/2016/dh/DH09.pdf	
Evaluating the re-identification risk of a clinical study report anonymized under EMA Policy 0070 and Health Canada Regulations	https://trialsjournal.biomedcentral.com/track/pdf/10.1186/s13063-020-4120-y.pdf	Re-identification risk
https://zenodo.org/communities/cadre/	https://zenodo.org/communities/cadre/	Australian framework for data privacy
Estimating the success of re-identifications in incomplete datasets using generative models	https://www.nature.com/articles/s41467-019-10933-3	Identifying that most data sets can be re-identified
A Systematic Review of Re-Identification Attacks on Health Data	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3229505/	
The National Centre for Research Methods Anonymisation: theory and practice; Mark Elliot 3 part series	https://www.ncrm.ac.uk/resources/online/all/?id=20430	3 part youtube video, supporting material
The Anonymisation Decision-Making Framework 2nd Edition: Overview Mark Elliot,	https://msrbcel.files.wordpress.com/2020/11/adf-2nd-version-overview.pdf	

Elaine Mackey & Kieron O'Hara		
Doing Research Differently: Archiving & Sharing Qualitative Data in Studies of Childhood, Education and Youth	https://apo.org.au/sites/default/files/resource-files/2020-05/apo-nid303357_0.pdf	
Evaluating re-identification risks with respect to the HIPAA privacy rule	https://pubmed.ncbi.nlm.nih.gov/20190059/	
The Potential of Research Drawing on Clinical Free Text to Bring Benefits to Patients in the United Kingdom: A Systematic Review of the Literature	https://www.frontiersin.org/articles/10.3389/fdgth.2021.606599/full	Systematic Review Article from Frontiers Health
Audit of District Court opinions	https://public.resource.org/scribd/7512583.pdf	
Experimentation with Personal Identifiable Information	https://www.scirp.org/pdf/IIM20120400004_52742027.pdf	
Privacy Detective: Detecting Private Information and Collective Privacy Behaviour in a Large Social Network	https://www.researchgate.net/publication/287317626_Privacy_Detective_Detecting_Private_Information_and_Collective_Privacy_Behavior_in_a_Large_Social_Network	
Experimentation with Personal Identifiable Information	https://www.scirp.org/html/5-8701169_21408.htm	
Practical Implications of Sharing Data.	https://www.lexjansen.com/pharmasug/2016/IB/PharmaSUG-2016-IB06.pdf	A Primer on Data Privacy, Anonymization, and De-Identification
Recommendations for Regulating Non-identifiable Data - Khalel El Eman	https://www.replica-analytics.com/web/default/files/Resources/Knowledgebase/10%20recommendations%204%20deid%20-	

	%20v11.pdf	
Privacy considerations in large language models	https://ai.googleblog.com/2020/12/privacy-considerations-in-large.html	
Coordinated Access for Data, Research and Environments (CADRE) – A Five Safes Implementation Framework for Sensitive Data in Humanities, Arts, and Social Sciences in Australia.	https://cadre5safes.org.au/about/	A system to:Increase the speed at which social sciences and related disciplines get access to sensitive data by means of the development of a shared and distributed sensitive data management platform using the Five Safes framework
Broken promises of Privacy: Responding to the surprising failure of anonymization	http://www.datascienceassn.org/sites/default/files/Broken%20Promises%20of%20Privacy%20-%20Responding%20to%20the%20Surprising%20Failure%20of%20Anonymization.pdf	
Statistical Confidentiality: Principles and Practice	https://link.springer.com/book/10.1007/978-1-4419-7802-8	
Disclosure Control of Microdata	https://www.jstor.org/stable/pdf/2289523.pdf?casa_token=im-RBiUewvQAAAAA:e_SZq54Ti9-amQQFMtPoK3bs6OiE4PpjJmN4Lsy1gIRb94fa_dP0QOEdZ_LzahNeIWYecofinCr8sFUWc-IgthaEtsvkL5ir1gwMkLIvDIf0iPXGA	
The FAIR Guiding Principles for scientific data management and stewardship	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/	