Exploring Pharmaceutical Industry Professionals' Views and Experiences of Implementing Anonymisation for Data Sharing: A Qualitative Interview Study

Nastazja Monika Graff[1†], Alex Hughes[2], Mark Elliot[1]

[1]The University of Manchester Cathie Marsh Institute, Manchester M15, United Kingdom of Great Britain and Northern Ireland
[2]Roche Products Ltd, Product Development Data Sciences, Hexagon Place , Welwyn Garden City, AL7 1TW, United Kingdom of Great Britain and Northern Ireland

Keywords: Data anonymisation; Data sharing; Clinical trials; Clinical study report; Expert interviews; Pharmaceutical industry

Citation: Graff N.M., Hughes A., Elliot M.: Exploring Pharmaceutical Industry Professionals' Views and Experiences of Implementing Anonymisation for Data Sharing: A Qualitative Interview Study. Data Intelligence, Vol. 7, Art. No.: 2025r27, pp. 723–744, 2025. DOI: https://doi.org/10.3724/2096-7004.di.2025.0059

ABSTRACT

Clinical trial data including information about the health and side effects experienced by participants has typically been held securely behind the walls of pharmaceutical companies. It is only in the last decade that the concept of anonymisation for data sharing has evolved within the industry. To characterise the obstacles that need to be overcome in this area, this study reports on professionals' views and experiences of anonymisation for data sharing in the pharmaceutical industry. Semi-structured qualitative interviews were carried out with professionals working in technical anonymisation and/or data sharing management roles at major pharmaceutical companies, as well as with a contract research organisation specialising in data anonymisation for the pharmaceutical industry. Thematic analysis of the interviews to define processes and characterise barriers and facilitators relevant to anonymisation for data sharing was conducted. Fourteen participants were interviewed from seven companies. Results identified two themes relating to challenges: (1) developing a standardised methodology and (2) limited resources and data science capabilities. Results also identified benefits of data sharing policies and opportunities for improvement of automated processes. This study demonstrates challenges and opportunities for the development of standardised and automated approaches to quantitative risk assessment that are tailored to CSRs.

[†] Corresponding author: Nastazja Monika Graff (E-mail: nastazja.laskowski@postgrad.manchester.ac.uk; ORCID: 0009-0008-9121-7788).

1. INTRODUCTION

Anonymisation processes for data sharing are a relatively recent priority for the pharmaceutical industry. In the last decade, companies have committed to delivering on principles of clinical trial transparency as set out in the joint EFPIA-PhRMA Principles for Responsible Clinical Trial Data Sharing, European Medicines Agency (EMA) Policy 0070 and other related policies [1-3]. The issues of how anonymisation should work in practice with clinical data, in manner which maintains sufficient utility for the purposes of data sharing is not well understood. It is noteworthy that, as of 2025, there is no standardised operating procedure for anonymisation of clinical trial datasets. Companies are consequently developing their own interpretations of regulatory data sharing policies and their own methodologies for delivering sufficiently de-identified data. The following quote from a pharmaceutical industry professional working in anonymisation for data sharing conveys the nuances and challenges that are yet to be overcome,

> *"On the vendor side, it's a tremendous amount of work for them to go through everything. That's why we really want to get to a way where it is more reproducible or to a point where we can treat different phases of studies differently. If people are more likely to look at phase III information, rather than some of the earlier phases or integrated data, treat that differently than single study data. We haven't come upon any kind of magic formula."* (P13[①])

Patient data has historically been safeguarded to avoid violations of clinical trial participants' human right to privacy and in accordance with confidentiality assurances given to patients (through participant information sheets). The first step towards implementing anonymisation for data sharing was taken when individual companies started to open their doors to researchers in 2010–2015 through policies including the 'Roche Global Policy on Sharing of Clinical Study Information' [4]. The Roche data sharing policy sets out principles that are characteristic of most data sharing policies in the pharmaceutical industry. Under the policy, researchers engaging in rigorous, independent scientific research can access patient-level data through the cross-industry request site 'vivli.org'. It requires researchers to submit a research proposal that proves scientific merit, which is assessed by an Independent Review Panel managed by Vivli. Researchers sign a contract to protect against attempts to re-identify participants in the dataset, thereby providing protective measures towards patient privacy. A further step was taken in 2015, when EMA Policy 0070 (published in 2014) came into force. Alongside their usual clinical trial data submissions, pharmaceutical companies are required by the EMA since 2015 to submit an anonymised version of a clinical study report (CSR) for public release.

The goal of EMA Policy 0070 was to increase transparency in the pharmaceutical industry, and to encourage the use of a *quantitative* threshold for risks of re-identification associated with a CSR. Prior to Policy 0070, it was commonplace to carry out *qualitative* risk assessments involving manual review of the CSR and to assign a low/medium/high score for 'riskiness', which were later regarded as subjective and therefore unreliable by the EMA. Policy 0070 requires anonymised CSRs to meet a numerical threshold for the risk of re-identification, which should either be 0.09 or another number

[①] In this notation, 'P' refers to 'participant'. Here 'P13' means 'Participant 13' from the interviews.

accompanied by a justification[2] [2]. The specific value of 0.09 will be further described and discussed in this paper, as historical reasons for its application in the pharmaceutical industry are not clearly understood. The reference used by the EMA discusses some precedents upon which the choice of 0.09 was based [6]. Although the EMA require publication of CSRs, responsibility for GDPR compliance lies with the pharmaceutical companies. Pharmaceutical companies must be careful to meet the prescribed anonymisation standards, else they risk large fines from data protection authorities if a clinical trial participant is re-identified from the publicly available CSR. Fines vary depending on the jurisdiction and the specific circumstances of the breach.

This paper will explore the experience and views of professionals working in anonymisation and data sharing of clinical trial data on the practical implementation of quantitative risk assessments as part of anonymisation for data sharing under EMA Policy 0070 and related policies[3]. The implementation of these practices has not been straight forward. Despite being enforced in 2015, Policy 0070 was paused for several years due to the disruption caused by the move of the EMA headquarters from London to Amsterdam and then COVID-19. Consequently, pharmaceutical companies have only had a few years of practical experience submitting anonymised CSRs to the EMA and also have been dealing with the increased workload arising from these practices.

As introduced earlier in this section, Policy 0070 sets a quantitative threshold for the risk of re-identification associated with a CSR. As standard, all direct identifiers are removed from a CSR (e.g. names, hospital numbers, social security numbers) before sharing the data with researchers. Risks assessments are therefore based on 'indirect identifiers' which are commonly present across clinical trial data (e.g., age, race, gender, location). To reduce the likelihood of identifying someone in the dataset, the indirect identifiers are generalised to coarser categories or redacted so that participants blend into larger groups sharing the same values on the indirect identifier variables. The threshold of 0.09 set by Policy 0070 dictates that any single participant in the dataset should be 1 of at least 11 participants sharing the same values for the indirect identifiers. This concept of 'hiding in a crowd' is referred to as k-anonymity. When applying k-anonymity to a dataset, attributes are suppressed or generalised until each row is identical with at least k – 1 other rows [7]. The way in which k-anonymity is used to calculate the risk of re-identification is as follows:

$$\frac{1}{k} = x$$

- k is the number of participants in a group sharing the same values for the indirect identifiers.
- x is the estimated risk of re-identification.

[2] This criterion is lifted from EMA Policy 0700, referring to a calculation based on k-anonymity. K-anonymity is achieved when each of the released records becomes indistinguishable from at least k □ 1 other records [5].

[3] Here I refer to Health Canada Public Release of Clinical Information (PRCI), since the same CSR can be submitted for this and EMA Policy 0070. Both require adherence to the 0.09 threshold [3].

For professionals working in anonymisation, this equation underscores the importance of ensuring sufficiently large equivalence classes (groups of individuals with the same indirect identifiers) to minimise the risk of re-identification.

The iterative process of assessing and transforming a dataset to reduce the risk of re-identification of individual population units and/or the disclosure of information about individual population units through statistical processes including k-anonymity is called statistical disclosure control (SDC) [8]. It has been proposed that k-anonymity can be applied to large datasets to prevent re-identification of individuals [9]. The impact of meeting k = 11 threshold on data utility of the dataset often depends on the size of the data. Unlike large census datasets, CSRs contain relatively small numbers of participants (sometimes less than 1000). There is very little published literature about professionals' views on the implementation of quantitative risk thresholds for anonymisation and the challenges of data sharing in the pharmaceutical industry. This paper aims to fill the gap in the pharmaceutical and data science academic literature by reporting on the perspectives of professionals from the pharmaceutical industry working in data anonymisation and sharing.

2. RESEARCH QUESTIONS

Following from the foregoing, this study aims to answer the research questions below (the interview topic guide is found in Appendix 1),

1. What are views and experiences of professionals with data sharing and/or anonymisation responsibilities on the practical implementation of quantitative risk assessment during anonymisation of CSRs for data sharing?
2. What are the barriers and facilitators involved in quantifying the risks of re-identification and anonymising clinical study reports for data sharing in the pharmaceutical industry?

Subsidiary research questions that will help to understand the context for anonymisation practices are as follows:

3. What is the history, purpose and composition of departments involved in anonymisation and/or data sharing?
4. What are the steps involved in anonymisation of clinical study data?
5. What are the views of professionals on the future of anonymisation and/or data sharing?

Overall, these research questions progress from current practices to practical implementation and finally to future outlook. This approach allows for a thorough exploration of anonymisation practices, addressing current and future challenges. Research question 1 aims to explore the practical implementation of quantitative risk assessment since the introduction of policies such as EMA Policy 0070. This is important because quantitative anonymisation has become increasingly important in the pharmaceutical industry, offering a more data-driven approach to protecting patient privacy while maintaining data utility. Research question 2 seeks to identify barriers and facilitators in quantifying re-identification risks in CSRs. This follows from the first question, as understanding these factors can help to improve the implementation of quantitative risk assessment methods. Research question 3 provides context to the current anonymisation and data sharing processes. This background information is essential for understanding the organisational structure and evolution of anonymisation practices. For example, whether the methodology is being driven by statisticians or computer scientists. Research questions 4 follws from the previous questions, as it provides a practical understanding of the anonymisation process, which is crucial for identifying potential areas of improvement. Research question 5 involves a forward-looking perspective for understanding potential trends and challenges in the field, which can inform future research and practices.

3. METHODOLOGY

3.1 Study design

As described above, the aim of this study is to gather, collate and interpret information about professionals' views and experiences of the practical implementation of quantitative risk assessment during anonymisation of CSRs for data sharing in the pharmaceutical industry. The outputs of this research are intended to provide guidance for practitioners and academics interested in developing their methodology to meet data transparency obligations.

The state of knowledge regarding the application of SDC and anonymisation to CSRs in the pharmaceutical industry is pre-theoretical, therefore the research process adopted in this study is necessarily inductive. This inductive approach allows themes and patterns to emerge organically from the raw data, rather than testing pre-existing theories [10]. Semi-structured interviews were chosen to capture rich data about the views and experiences of pharmaceutical industry professionals working in data sharing and anonymisation. The semi-structured format is particularly effective for exploring in-depth information and evidence while maintaining flexibility and adaptability, making it particularly effective for eliciting contextual insights in complex research settings [11]. Thematic analysis was applied to the interview transcripts following the approach described in Braun & Clarke [12], and described in the 'Data Analysis' section of this study.

3.2 Target sample

Pharmaceutical industry professionals working in anonymisation and/or data sharing were the target population for this research study as both groups are stakeholders in the use of anonymisation for the purposes of data sharing. One of the authors, Alex Hughes (AH) acted as a facilitator for the recruitment of pharmaceutical industry professionals into the study due to his own employment in the pharmaceutical industry and network of connections. Inclusion criteria for semi-structured interviews were: English speaking and employed within an anonymisation or data sharing department at a pharmaceutical company. There were no exclusion criteria. In particular, there were no restrictions on number of years spent in anonymisation or data sharing or the size of the employing company.

The aim was to conduct ten interviews with industry professionals from at least five different companies. Purposive sampling and snowball sampling were used by Alex Hughes (AH) and Nastazja Monika Graff (NMG) who approached professionals that were employed in an appropriate role or department to ask them to take part in the interview or to recommend a colleague that could take part. Purposive sampling was chosen to ensure the inclusion of information-rich cases, specifically professionals with expertise in the implementation and logistics of anonymisation in the pharmaceutical industry [13]. Snowball sampling was used as a complimentary method to purposive sampling, allowing us to expand our participant pool by leveraging professional networks [14]. This approach was especially useful given the specialised nature of anonymisation work in the pharmaceutical industry, facilitating access to additional experts and maintaining a focus on individuals with relevant expertise. This approach aligned with similar studies in the pharmaceutical industry, such as those exploring industry professionals' beliefs about patient and public involvement in medicines research and development [15].

3.3 Procedures

3.3.1 Participant recruitment

Participants were approached by AH and NMG via email. The email enclosed an interview participant information pack. Contact details for NMG were included at the end of the email so the participants could contact NMG directly to express interest, ask questions and return the consent form. Consent was recorded via written or digital signature on the consent form and returned via email. Online interviews were organised at the participants convenience. Participants were welcomed to view the topic guide enclosed in the participant information pack before agreeing to participate in the study to assess whether they felt comfortable contributing (Appendix 1).

3.3.2 Data collection and topic guide

The topic guide was created to ensure each interview addressed the research aims in this study (Appendix 1). The first section of questions elicited the participant's job role and the history of the anonymisation or data sharing department that the participant worked in. The second section of questions covered the technical procedures involved in assessing risks of re-identification for datasets, anonymising datasets and sharing them. The final section contained open questions about barriers, facilitators and the future of data sharing/anonymisation. The topic guide provided a set of open-ended prompts for topics of discussion, whilst allowing room for conversational exchange and for participants to expand on issues of interest to them.

Interviews were conducted between February and June 2023. Interviews were carried out via video-call and the audio stream was recorded for transcription. Recording of video was avoided to minimise intrusiveness and to encourage participants to speak freely. The interviewer notified the participant that the interview was being audio recorded before asking any questions.

3.3.3 Data treatment and storage

Audio recordings of interviews were uploaded to a secure server at the University of Manchester and deleted from the audio recording device. Files were uploaded to Otter AI for automated transcription. Transcriptions were also stored on the secure server. The data will remain on the servers for up to three years to allow for further analysis and review. To protect the confidentiality of the participants, transcripts were pseudonymised by replacing participant names with a unique identifier and storing the list of names and identifiers in a separate password-protected file on a secure server at the University of Manchester.

3.4 Data Analysis

Braun and Clarke's iterative process of thematic analysis was used for the analysis of interview transcripts. This is an inductive method consisting of (1) becoming familiar with the interview data, (2) generating codes from the data, (3) generating themes from the codes, (4) reviewing themes, (5) defining and naming core themes and (6) locating exemplar quotes to illustrate the themes [12]. In practice, this involved reading the transcripts, highlighting interesting pieces of information that addressed the research questions and assigning these snippets a description (code), comparing all the codes and combining those that reoccur and overlap into themes.

The interviews were also analysed for the steps involved in anonymisation to create a diagram of the underpinning anonymisation process collated across all interviews. This involved highlighting any procedural descriptions and collating them together. Steps which were common across companies were compiled into the diagram. Small variations in methodology were omitted for purposes of clarity (Figure 1).

4. RESULTS

4.1 Sample demographics

Fourteen pharmaceutical industry professionals participated in interviews lasting between thirty and sixty minutes with two of the authors (Table 1). Seven companies were represented in the interviewee list. There was an even split of participants working in data sharing and technical data anonymisation. Participants working on technical anonymisation tended to come from a statistical programming background, whereas participants working in data sharing came from a mixture of backgrounds. Years of experience in the pharmaceutical industry ranged from ten to thirty years, whereas years of experience in anonymisation and/or data sharing ranged from five to ten years (values rounded to multiples of 5 for participant anonymity) in line with the emergence of departments with specific anonymisation responsibilities in 2015.

4.2 The anonymisation process of CSRs

Pharmaceutical industry CSRs reach the team responsible for anonymisation for data sharing with personal protected information (PPI) intact. The team is then responsible for the iterative process of

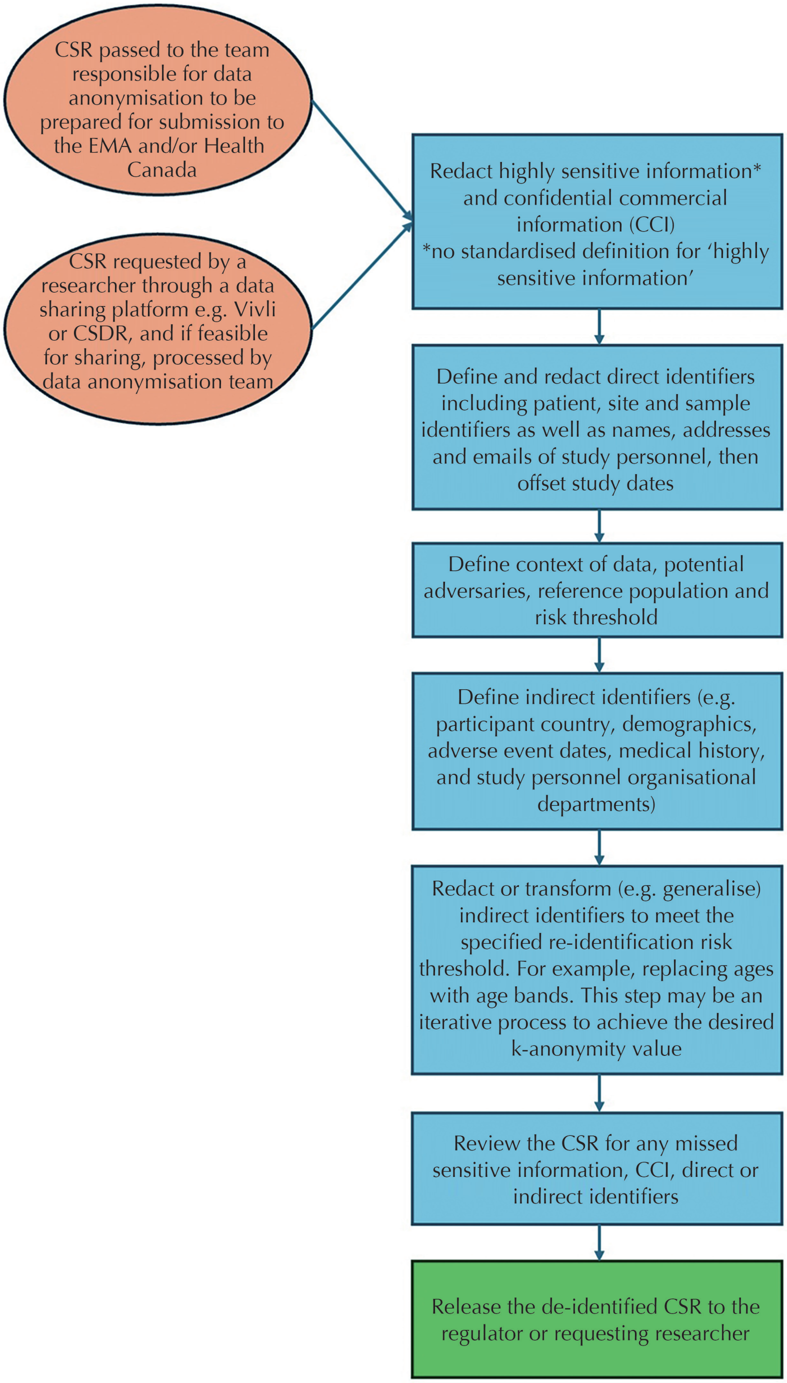
CSR passed to the team responsible for data anonymisation to be prepared for submission to the EMA and/or Health Canada
CSR requested by a researcher through a data sharing platform e.g. Vivli or CSDR, and if feasible for sharing, processed by data anonymisation team
Redact highly sensitive information* and confidential commercial information (CCI)
*no standardised definition for 'highly sensitive information'
Define and redact direct identifiers including patient, site and sample identifiers as well as names, addresses and emails of study personnel, then offset study dates
Define context of data, potential adversaries, reference population and risk threshold
Define indirect identifiers (e.g. participant country, demographics, adverse event dates, medical history, and study personnel organisational departments)
Redact or transform (e.g. generalise) indirect identifiers to meet the specified re-identification risk threshold. For example, replacing ages with age bands. This step may be an iterative process to achieve the desired k-anonymity value
Review the CSR for any missed sensitive information, CCI, direct or indirect identifiers
Release the de-identified CSR to the regulator or requesting researcher

Figure 1. CSR Anonymisation Process showing common steps across companies.

Table 1. Summary of participant information.

Interview ID	Area of Expertise	Years in Data Sharing and/ or Anonymisation (rounded to multiple of 5)	Years in Pharmaceutical Industry Overall (rounded to multiple of 5)
P1	Clinical Trial Disclosure, Data Sharing	10	20
P2	Data Sharing	10	10
P3	Data Sharing	Not disclosed	20
P4	Clinical trial disclosure, Data Sharing, Programming	5	30
P5	Data Sharing	10	30
P6	Data Anonymisation, Statistics	5	20
P7	Statistics	5	15
P8	Data Sharing	5	Not disclosed
P9	Clinical Trial Disclosure	Not disclosed	Not disclosed
P10	Data Anonymisation, Statistics	10	20
P11	Data Anonymisation	10	30
P12	Data Anonymisation, Clinical Trial Disclosure	10	30
P13	Statistics, Programming	5	30
P14	Clinical Trial Disclosure, Data Sharing	10	15

anonymising CSRs for release either into the public domain under EMA Policy 0070 or for sharing with researchers. The threshold for risk of re-identification can be different for researchers and it is assessed on a case-by-case basis. The 'data situation' is taken into account including who is handling the data and where it is stored [16]. Information contained in the CSRs has varying levels of disclosiveness and/or sensitivity and this drives decisions regarding whether to redact the information or to modify it using methods such as generalisation. Figure 1 provides an outline of the major steps in the anonymisation process.

4.3 Themes

Thematic analysis of the interview transcripts identified three themes arising from the interview transcripts about the views and experiences of professionals. The process of thematic analysis was based on the method of Braun & Clarke described in Section 3.4. The three themes identified were challenges to developing a standardised methodology, challenges relating to resources and data science capabilities, and benefits of data sharing policies. Each theme, along with relevant sub-themes is illustrated below (Figure 2).

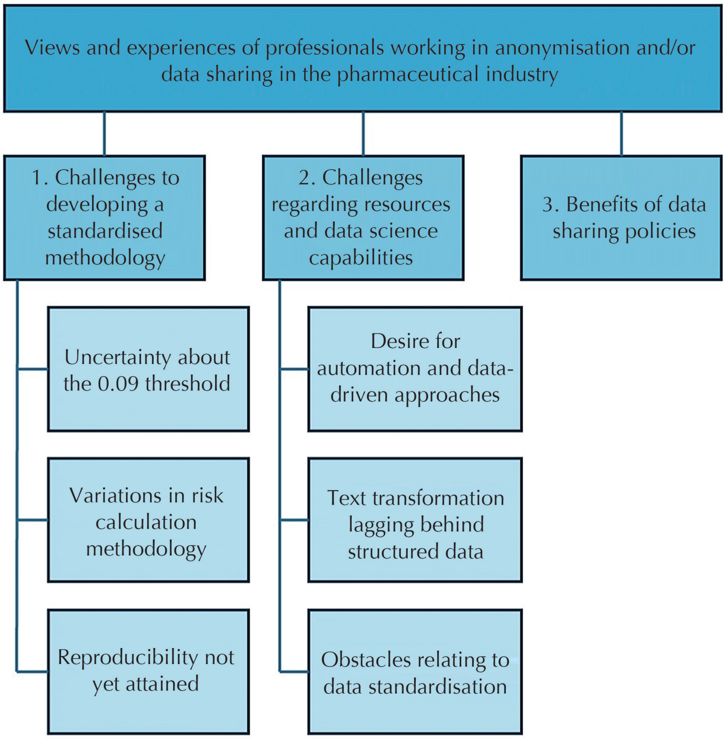
Views and experiences of professionals working in anonymisation and/or data sharing in the pharmaceutical industry
1. Challenges to developing a standardised methodology
2. Challenges regarding resources and data science capabilities
3. Benefits of data sharing policies
Uncertainty about the 0.09 threshold
Variations in risk calculation methodology
Reproducibility not yet attained
Desire for automation and data-driven approaches
Text transformation lagging behind structured data
Obstacles relating to data standardisation

Figure 2. The thematic analysis themes and sub-themes.

Pharmaceutical industry professionals described interconnected challenges surrounding the introduction of a quantitative re-identification threshold within anonymisation for data sharing. Participants expressed a lack of clarity surrounding the historical reasons for the 0.09 threshold, reported variable approaches to achieving k-anonymity and unpredictability in the feedback from regulatory agencies. The resource-intensive nature of anonymisation for data sharing has impacted teams and highlighted challenges relating to data science capabilities, particularly text transformation, and the need for automation. However, not all the data was relating to challenges, and participants also acknowledged the benefits of data sharing, particularly the improved relationship with researchers.

4.3.1 Theme 1: Challenges to developing a standardised methodology

This theme captures the challenges behind developing a standardised methodology for anonymisation for data sharing in the pharmaceutical industry. Participants referred to this as a 'gold standard methodology' that could be used across companies. This would be a collective agreed standard and a common methodology shared throughout the community, most likely via PHUSE (Pharmaceutical Users Software Exchange) found at the website address 'phuse.global'. Cross-company working was not seen as a barrier and participants felt companies were working together to solve a difficult problem. There was no perceived competitive advantage to solving this problem on the part of the pharmaceutical companies, whereas there was some on the vendor side. The pharmaceutical companies believed they are working together to improve public trust. The participants described uncertainties surrounding the appropriateness of the 0.09 threshold in the context of clinical trial data due to its impact on data utility. Variations in the detail of risk calculation methods and the reference population further contribute to the complexity. Additionally, lack of regularity in the content of the feedback from regulatory agencies had made it difficult to meet expectations for a reproducible approach.

1a) Uncertainties about the 0.09 threshold.

Limited understanding of the historical reasons of the 0.09 threshold for risks of re-identification were reported. Participants used the threshold because the EMA guidance stipulates it. One participant expressed that the exact value of the threshold was not interesting to them, but rather just the existence of an objective measure to adhere to whilst anonymising was very helpful compared to before the introduction of this threshold.

Participants were keen to access original sources and literature that provides the rationale for this threshold but reported difficulties in finding these. Questions arose around whether a different threshold would be more suitable for clinical studies due to their small sample sizes, especially in Phases I and II, and rare disease trials.

Six participants said the 0.09 threshold for risk of re-identification had a significant effect on data utility in clinical trial datasets which contained small numbers of participants. The following quotes illustrate professionals' views on the 0.09 threshold:

P10: *"We need to build groups of 11 to get anywhere… instinctively after seven or eight years of looking into these things… I think this is too conservative".*

P9: *"This space is about 10 years old… you start out conservative and realise maybe some of these rules can be relaxed a little bit because they're not working for us".*

P6: *"[In rare diseases] you can't really move to a quantitative approach with serious adverse events because we will hardly be able to see anything".*

P3: *"We don't want to give [researchers] crappy data. So if we could enhance utility and protect patient privacy, we would do it in a heartbeat".*

1b) Variations in risk calculation methodology.

From the interview data the authors of this paper could not define a clear consensus across companies on the exact process to achieve k-anonymity. Particularly the reference population used, variables selected for SDC and definition of 'sensitive variables' were all reported differently and to differing levels of detail. Possible causes for this will be discussed in the 'Discussion' section. For example, the reference population was used to calculate k-anonymity in three different ways:

- the participants with patient-level data in the CSR (i.e. those reported in the patient narratives),
- all participants contributing to the clinical trial dataset including aggregate data (i.e. including those not described in the patient narratives),
- or a combination of all the populations of studies within the submission/disease area.

Whilst some participants referred to the use of average risk across the dataset, other participants referred to the use of maximum risk present in the dataset. A hybrid approach was also described, where average risk was used but the data was reviewed for any unique records that require transformation (e.g. unique ages being banded). It was not clear by the end of this study what to attribute variation in reported methodology to. On reflection, it could be attributed to participants preparing the CSRs for different purposes (e.g., Health Insurance Portability and Accountability Act compliant versus General Data Protection Regulation compliant), as this is a plausible explanation.

There was variation in how participants approached free–text patient narratives. Some companies redacted all free-text information apart from the adverse event term, whereas others only redacted information classified as highly sensitive from the adverse event patient narratives after SDC was carried out on the indirect identifiers. There was no standardised definition for 'sensitive information' across companies and participants said that departments defined this internally.

1c) Consistency not yet attained.

Feedback from regulatory agencies was reported to be inconsistent by several of the participants[④]. 'Inconsistent' here refers to receiving different comments back on whether the anonymisation approach is acceptable upon each submission despite applying the same approach across multiple submissions (with similar characteristics e.g. sample size). The participants reported feeling that both they and the regulators were still on a learning curve of deciding how to appropriately transform data to meet the risk of re-identification threshold. Observations of inconsistency in feedback are captured by the following quotes:

P14: *"We've gotten conflicting feedback regarding which approach to take. It's been challenging to meet the expectations and take learnings into the next time we're doing a package, we feel like we're starting over with each one."*

P11: *"It seems that each package is being reviewed by a different person who either hasn't been properly trained or that has different viewpoints from others. So it seems like the authorities don't really follow or understand."*

[④] Interviewees used the word 'reproducibility' when talking about a desire to get a consistent response back from regulators when applying their anonymisation principles to studies. Since the response received has been inconsistent, we use this term in our results.

There was a desire to develop methods that could be applied to multiple submissions packages. Suggestions included developing methods that were specific to the clinical trial type due to the different participant numbers (i.e., Phase I, II or III). Some participants also expressed a more general desire to streamline the process by developing basic principles that could be followed regardless of the submission, population or drug.

Despite tensions between companies and regulatory agencies, companies demonstrated an understanding that their relationship with the regulators is mutually beneficial, as illustrated by the following quote:

P4: *"[Regulators are] pushing companies to do better in this… you've got to turn the heat up incrementally… it creates this dynamic tension".*

4.3.2 Theme 2: Challenges regarding resources and data science capabilities

Theme 2 captures participants' comments on challenges relating to internal resources and companies' current data science capabilities. Given the increased workload of preparing an anonymised CSR for every clinical trial, there was a reported desire for automation and data-driven approaches to anonymisation. Some of the processes were outsourced to third-party vendors, but even here there were challenges relating to natural language processing (NLP) being insufficiently developed, with the state-of-the-art in transformation of text lagging behind that of transformation of structured data.

2a) Desire for automation and data-driven approaches.

The strain of the resource-intensive nature of anonymisation and data sharing on internal resources is unsurprising considering that teams with anonymisation and data sharing responsibilities within pharmaceutical companies were reported to be small (i.e. teams of less than five professionals). Since the introduction of policies necessitating the submission of an anonymised version of the CSR to the EMA and other regulatory agencies, pharmaceutical companies have experienced a dramatically increased workload.

P13: *"I'm really wondering sometimes how academic researchers can even think about complying with some of the requirements. Because this takes a lot of work, to fully comply and make sure you follow all the rules under the GDPR for data privacy."*

Most participants reported developing principles for anonymising datasets in-house by utilising the programming backgrounds of those working in the anonymisation team. Some participants employed the help of an external company specialising in anonymisation to review, transform and redact CSRs. This would typically be done in line with guidelines developed in-house by the pharmaceutical company.

P10: *"I oversee what's going on…check the rules for anonymisation and redaction… but I don't go with the black marker and redact those things myself… it's a nightmare to go through 1000s of pages".*

P12: *"We [pharmaceutical professionals] don't actually do the technical stuff, it's a bit of a black box to us. We send the data over, we give specific requirements, we look what we get back and if needs be we ask for a rework"*.

For those participants who carried out review and anonymisation of CSRs in-house, some utilised the structured data from which CSRs originate as a dictionary for direct and indirect identifiers to search for in the CSR. Furthermore, a first pass look at this structured data could help form a judgement on which variables may need to be redacted and which transformed.

Participants expressed a desire to use data-driven approaches to decrease the use of redactions and move towards the use of transformation and text replacement to preserve data utility (e.g. mapping medical terms to higher level terms). There was a feeling that technologies which aid automation, including NLP, would become very important soon but their exact role was still unclear, as illustrated in the following quote:

P4: *"I try to listen closely about what people think about it [NLP]… it's such a rapidly evolving environment, where it's going is very difficult to predict… it's a bit like a tsunami, you don't know what the impact of the wave is but you better get yourself moving"*.

2b) Text transformation lagging behind that of structured data.

Whilst data transformation of structured numerical data has been gaining sophistication, transformations of textual data have been lagging behind. Since the introduction of text replacement instead of black box redactions, text replacement boxes have been noticeable due to basic issues with font and sizing. Participants reported knowledge that external companies were attempting to improve the restructuring of CSRs (usually provided in PDF format) to accommodate text replacement boxes that adapt to the new character length of the replacement word, however, at the time of writing this paper this is still a work in progress. Generally, most of the companies were not yet using NLP and processing of unstructured text data was either rule based or through manual review of all the text by a human.

P8: *"[The external companies helping us] spend a lot of time manually doing this [manual review of CSRs] and I know it can be done better"*.

P7: *"Even if we have NLP… still there would be a need to review it manually. We have been using NLP more for names, addresses or signatures but not for the adverse events or sensitive information"*.

2c) Obstacles relating to data standardisation and harmonisation.

The interview data showed that when it comes to data anonymisation and sharing, both structured and unstructured data would benefit from data standardisation. Discrepancies in data formatting were reported to create obstacles in automating the anonymisation process. The major issue being that if the data is not standardised, algorithms will struggle to detect and transform identifiers, leading to increased need for human review and strain on resources.

P4: *"There is a high degree of variance even within the structured data… if you can't even get your structured data to be consistent across your studies, what does it look and feel like for your unstructured piece?"*

P3: *"Can we improve the hours of work that people have to do upfront with data harmonisation without jeopardising regulatory filings because the FDA and EMA have their own peculiarities about how data has to be organised?"*

Furthermore, within the unstructured data discrepancies are sometimes as simple as differences in spelling. This highlights the simplicity of this obstacle, and yet the need for resource-intensive attention to detail to harmonise it retrospectively.

P6: *"In our company spelling in the CSR is different than data for the preferred terms. English spelling is part of MedDRA and [we get] American from the medical writers"*.

P5: *"If people took standards more seriously, then our job becomes easier. If we get clean data, and people explain exactly what it is"*.

Data formatting discrepancies also create obstacles for data sharing since reformatting the data for a research partner requires a large investment of resources.

P9: *"I think we've learnt we'll say no to [data reformatting for researchers], because it really took up a lot of resources from other projects"*.

4.3.3 Theme 3: Benefits of data sharing policies

Despite challenges in the implementation of a quantitative risk assessment threshold during anonymisation for data sharing, participants discussed benefits to the new approach as well. This theme encompasses the benefits of data sharing policies, which came up often but variations were limited therefore this section is brief. The length of the section does not reflect the importance, as this was a prominent theme in the data.

By removing the objectivity of a qualitative risk assessment, it has been possible to streamline the process and move forward with more certainty in data sharing. This benefit is illustrated accurately by the following quote:

P3: *"Now with an objective score calculation, we can do it [anonymisation] quicker and say 'that's the number and we're good'"*.

Since the introduction of data sharing policies, participants reported seeing a steady and continuous increase in requests for clinical trial data from researchers. This suggests that implementing anonymisation for data sharing has succeeded in promoting a positive relationship between the pharmaceutical industry and researchers. Participants felt that by sharing anonymised CSRs, they were contributing towards external research and collaboration. Whilst recognising the progress made, participants were also conscious of the existing obstacles to data sharing.

P8: *"Now that the industry has committed itself to making the data available and has a transparent process with a neutral platform and arbiters in place, it [accessing data] is easier than it was 10 years ago. But there's still a high level of competence and scientific understanding to do it"*.

5. DISCUSSION

This study investigated the views and experiences of pharmaceutical industry professionals involved in the implementation of anonymisation for data sharing, with a particular focus on quantitative assessments for the risks of re-identification. Thematic analysis of semi-structured interviews with fourteen professionals working in anonymisation and/or data sharing enabled us to collate and interpret insights about the practical implementation of data transparency policies. Three primary themes emerged: challenges to developing a standardised methodology, challenges relating to resources and data science capabilities, and benefits of data sharing policies. We found that implementing data transparency policies is a resource-intensive process, which requires progress in automation and processing of unstructured data. The quantitative risk assessment process within anonymisation is complex, involving consideration of clinical trial dataset size and selection of appropriate SDC methodology. A standardised methodology is not yet defined and pharmaceutical companies have not yet reached a level which reproducibly satisfies the regulatory agencies.

The findings reported in this study support exploration of different re-identification risk thresholds. The threshold rule $n > 10$ is often used in output SDC, although not exclusively as other numbers are also used by various organisations [17]. Considering the comparatively small dataset size of a Phase I or II clinical trial, it would be in the interest of data utility to research the use of a smaller '*n*' for k-anonymity. We also consider that data anonymisation models other than k-anonymity may need to be explored for clinical trial data if sensitive attributes (e.g., adverse events and medical history) are disclosed. For example, l-diversity is an extension of k-anonymity that promotes intra-group diversity of sensitive values [18].

The findings further show that efforts should not stop at the point of selecting an appropriate methodology, the methodology must also be documented in detail and disseminated between pharmaceutical companies in a collaborative effort through forums such as PHUSE (Pharmaceutical Users Software Exchange) found at the website address 'phuse.global'. Efforts in this direction have been made by Lukasz Kniola of Biogen who has published in the PHUSE forum on 'Calculating the Risk of Re-Identification of Patient-Level Data Using Quantitative Approach', 'Text as Data in the Context of Anonymising Clinical Study Reports' and 'Data Anonymisation & Risk Assessment-Process Map and Automation Efforts' [19-21]. Furthermore, a collaborative PHUSE White Paper on 'Data Anonymisation and Risk Assessment' was produced in 2020 [22]. However, it is unclear to what extent the principles in the work of Kniola and in the White Paper are implemented across industry. A brief review of the Roche data sharing commitment on the data sharing platform 'Vivli', suggests that the Anonymization

Decision-making Framework [16], the Guide to the De-Identification of Personal Health Information [23], the PHUSE de-identification standards [24] and the Transcelerate standards [25] are the main principles followed by Roche. Despite a variety of principles on offer that could be adopted across industry, companies have not yet succeeded in fully standardising anonymisation process. Our study showed a lack of clarity and consensus within and between pharmaceutical companies on the best methodology for achieving the 0.09 threshold set by EMA Policy 0070. Further research and cross-industry publications in the style of those by Kniola would be needed to define and disseminate the methodologies used across the industry. Considering the inconsistency of responses participants received from the regulatory agencies when replicating their methodology in submissions, development of a standardised methodology would also benefit from a statement on preferred methodology to achieve the 0.09 threshold from the side of the regulators.

Automation was identified as a current challenge and future opportunity to facilitate the implementation of quantitative risk assessment during anonymisation for data sharing. A major obstacle to overcome here is processing of the unstructured data in adverse event patient narratives. Recent studies show that NLP can carry out detection of medical adverse events from free-text medical narratives [26-27]. Such studies tend to focus on a specific therapeutic area and hospital documents do not follow the same medical writing guidelines as adverse event patient narratives in CSRs. Therefore, to accurately assess current possibilities for NLP in CSRs, it would be important to characterise the data contained in different CSR patient narratives across different therapeutic areas. Following characterisation of the data, it may be possible to develop an appropriate NLP pipeline to automate the review of patient narratives. These make up a large portion of unstructured data in the CSR and would currently use a significant portion of human resources for manual review if they are not entirely redacted. Considering that the regulatory agencies are encouraging pharmaceutical companies to use techniques other than redaction, our recommendation is to assess the capabilities and move towards the use of NLP for CSRs. Clinical NLP in the UK has improved substantially in the last 15 years. For NLP, the total budget from funded grants in the UK in the period 2019–2022 was 80 times that of 2007–2010 [28] however, data standardisation remains an obstacle. The Clinical Data Interchange Standards Consortium (CDISC) have produced the Study Data Tabulation Model (SDTM) standard structure for human clinical trial data tabulations and the Analysis Data Model (ADaM) standards for subject-level analysis files and basic data structure for a variety of analysis methods. Whilst these standards have been widely adopted by the pharmaceutical industry, further research is needed to assess unstructured patient narratives for further irregularities such as use of English versus American spellings, as reported in this study.

Recognition of the benefits of implementing anonymisation for data sharing in the pharmaceutical industry is an important part of the findings in this study. Participants felt that they were contributing to a good cause and directly contributing to an improved perception of the pharmaceutical industry by facilitating the sharing of de-identified CSRs. Studies show that since the introduction of the EFPIA-PhRMA Principles for Responsible Clinical Trial Data Sharing, most of the top fifty pharmaceutical companies have indeed publicly committed to the disclosure of trial data [29]. Nevertheless, studies do show that there is still much room for improvement in terms of time taken to share data [30]. As discussed above, we feel that automation of anonymisation process, including that for unstructured data, has potential reduce the time and resources needed to deliver an anonymised CSR. Data owners can achieve cost and time savings through reduced compliance burdens and streamlined data management by adopting anonymisation and standardisation processes. These methods lower re-identification risks and manual handling expenses, while reputation gains from ethical sharing and participation in federated learning ecosystems create value beyond direct monetisation.

6. CONCLUSION

This study investigated the views and experiences of pharmaceutical industry professionals involved in the implementation of anonymisation for data sharing, with a particular focus on quantitative assessments for the risks of re-identification. We found that implementation of anonymisation for data sharing involves challenges to developing a standardised methodology and challenges to internal resources and data science capabilities. Despite these, participants perceived a benefit to the reputation of the pharmaceutical company amongst researchers and the public.

Looking forward, the work presented in this paper could be enhanced by applying quantitative textual analysis to complement the thematic analysis by identifying patterns and frequencies in the data. Grounded theory could also be used for more in-depth comparative analyses on themes across different participant groups. While these approaches extend beyond the scope of this study, they represent valuable directions for advancing the methodological rigour of qualitative research. More research is also needed into the application of SDC to clinical trial data specifically and possibilities surrounding NLP of unstructured patient narratives detailing adverse events.

ABBREVIATIONS

ADaM: Analysis Data Model.
AH: Alex Hughes (co-author).
CDISC: Clinical Data Interchange Standards Consortium.
CSR: Clinical Study Report.
EMA: European Medicines Agency.
NLP: Natural Language Processing.
NMG: Nastazja Monika Graff (author).
PPI: Personal Protected Information.
PHUSE: Pharmaceutical Users Software Exchange.
SDC: Statistical Disclosure Control.
SDTM: Study Data Tabulation Model.

AUTHOR CONTRIBUTIONS

Nastazja Monika Graff performed the research and collected and analysed the data. Alex Hughes contributed to designing the research framework and data collection. Mark Elliot contributed towards proposing the research problem and designing the research framework. The manuscript was written and revised by Nastazja Monika Graff and Mark Elliot.

ACKNOWLEDGMENTS

We would like to acknowledge Professor Goran Nenadic for his input on refining the research problem and designing the research framework.

REFERENCES

[1] PhRMA and EFPIA, “Principles for Responsible Data Sharing,” 2023. [Online]. Available: https://www.efpia.eu/media/qndlfduy/phrmaefpiaprinciplesforresponsibledatasharing2023.pdf. [Accessed: 22-Nov-2023].

[2] European Medicines Agency (EMA), “European Medicines Agency policy on publication of clinical data for medicinal products for human use,” 2014. [Online]. Available: https://www.ema.europa.eu/en/documents/other/european-medicines-agency-policy-publication-clinical-data-medicinal-products-human-use_en.pdf. [Accessed: 12-Feb-2024].

[3] Health Canada, “Guidance document-Public release of clinical information in drug submissions and medical device applications,” 2019. [Online]. Available: https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance/document.html. [Accessed: 15-Feb-2024].

[4] Roche, “Roche global policy on sharing of clinical study information (Version 2.1),” 2020. [Online]. Available: https://assets.cwp.roche.com/f/126832/x/091fe27314/roche-global-policy-on-sharing-of-clinical-study-informationv2-1-april2020-1.pdf. [Accessed: 28-Feb-2024].

[5] P. Samarati and L. Sweeney, “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression,” Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, Palo Alto, California, 1998.

[6] B. Lo, “Sharing clinical trial data: maximizing benefits, minimizing risk,” *JAMA*, vol. 313, no. 8, pp. 793–794, 2015.

[7] L. Sweeney, “Achieving k-anonymity privacy protection using generalization and suppression,” *Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 571–588, 2002.

[8] M.J. Elliot, “Statistical disclosure control,” in *Encyclopedia of Social Measurement*, K. Kempf-Leonard, Ed., Oxford: Elsevier/Academic Press, 2005, pp. 663–670.

[9] L. Sweeney, “k-anonymity: A model for protecting privacy,” *Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[10] D.R. Thomas, “A general inductive approach for qualitative data analysis,” 2003.

[11] R. Ruslin, S. Mashuri, M.S.A. Rasak, F. Alhabsyi, and H. Syam, “Semi-structured Interview: A methodological reflection on the development of a qualitative research instrument in educational studies,” *IOSR Journal of Research & Method in Education (IOSR-JRME)*, vol. 12, no. 1, pp. 22–29, 2022.

[12] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[13] M. Naderifar, H. Goli, and F. Ghaljaie, "Snowball sampling: A purposeful method of sampling in qualitative research," *Strides in Development of Medical Education*, vol. 14, no. 3, 2017.

[14] S.J. Stratton, "Purposeful sampling: advantages and pitfalls," *Prehospital and Disaster Medicine*, vol. 39, no. 2, pp. 121–122, 2024.

[15] S. Parsons, B. Starling, C. Mullan-Jensen, S.G. Tham, K. Warner, and K. Wever, "What do pharmaceutical industry professionals in Europe believe about involving patients and the public in research and development of medicines? A qualitative interview study," *BMJ Open*, vol. 6, no. 1, p. e008928, 2016.

[16] M. Elliot, E. Mackey, and K. O'Hara, *The anonymisation decision-making framework 2nd Edition: European practitioners' guide*, UKAN, Manchester, 2020.

[17] European Commission, "Guidelines for Output Checking in Data Without Borders," 2013. [Online]. Available: https://cros-legacy.ec.europa.eu/system/files/dwb_standalone-document_output-checking-guidelines.pdf. [Accessed: 22-Nov-2023].

[18] A. Machanavajjhala, et al., "l-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

[19] L. Kniola, "Calculating the Risk of Re-Identification of Patient-Level Data Using Quantitative Approach," PHUSE Annual Conference, 2016. [Online]. Available: https://lexjansen.com/phuse/2016/dh/DH09.pdf. [Accessed: 22-Nov-2023].

[20] L. Kniola, et al., "Text as Data in the Context of Anonymising Clinical Study Reports," PHUSE EU Connect 2018. [Online]. Available: https://www.lexjansen.com/phuse/2018/dh/DH04.pdf. [Accessed: 22-Nov-2023].

[21] L. Kniola, "Data Anonymisation & Risk Assessment-Process Map and Automation Efforts," PHUSE EU Connect 2019. [Online]. Available: https://www.lexjansen.com/phuse/2019/pp/PP24.pdf. [Accessed: 22-Nov-2023].

[22] PHUSE Data Transparency Working Group, "Data Anonymisation and Risk Assessment Automation," 2020. [Online]. Available: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Data+Anonymisation+and+Risk+Assessment+Automation.pdf. [Accessed: 22-Nov-2023].

[23] K. El Emam, *Guide to the de-identification of personal health information*, CRC Press, New York, 2013.

[24] J.-M. Ferran, "PhUSE De-Identification Working Group: Providing De-Identification Standards to CDISC Data Models," Paper DH01, 2015.

[25] Transcelerate Biopharma, "De-identification and Anonymization of Individual Patient Data in Clinical Studies: A Model Approach," 2016.

[26] A. Borjali, et al., "Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: A case study of detecting total hip replacement dislocation," *Comput. Biol. Med.*, vol. 129, pp. 104–140, 2021.

[27] S.S. Graham, et al., "Associations between aggregate NLP-extracted conflicts of interest and adverse events by drug product," *Stud. Health Technol. Informatics*, vol. 290, pp. 405–409, 2022.

[28] H. Wu, et al., "A survey on clinical natural language processing in the United Kingdom from 2007 to 2022," *NPJ Digit. Med.*, vol. 5, no. 1, pp. 1–15, 2022.

[29] S. Baronikova, et al., "Commitments by the biopharmaceutical industry to clinical trial transparency: the evolving environment," *BMJ Evid. Based Med.*, vol. 24, no. 5, pp. 177–184, 2019.

[30] M. Ventresca, et al., "Obtaining and managing data sets for individual participant data meta-analysis: scoping review and practical guide," *BMC Med. Res. Methodol.*, vol. 20, no. 1, pp. 1–18, 2020.

AUTHOR BIOGRAPHY

Nastazja Monika Graff is a PhD Candidate at the University of Manchester as part of the Centre for Doctoral Training in Data Analytics & Society. Her PhD topic is 'Data anonymisation and data privacy for clinical trial data' in collaboration with industry partner Roche pharmaceuticals.

APPENDIX 1 (INTERVIEW TOPIC GUIDE)

Specific questions about the participant and their department.

1. What is your job title?

2. What is your professional background?

3. How many people do you work with on data sharing/risks of re-identification?

4. What are the professional backgrounds of your colleagues?

5. When did the department form?

6. How did the department form?

7. What regulation or data sharing platform do you prepare CSRs for?

8. How many (if any) submissions have you made under Health Canada PRCI/EMA Policy 0070/EU CTR?

Specific questions about the anonymisation methods.

9. What was the old approach to anonymisation?

10. How has the approach changed?

11. How do you perform quantitative risk assessment (if any)?

12. Do you use a vendor for anonymisation?

13. How do you recommend handling small/rare disease datasets?

14. Do you tailor to bespoke requests from researchers (via Vivli)?

General questions about anonymisation under the new policies.

15. Are you involved in liaising with regulatory agencies?

16. What is your opinion on the 0.09 threshold?

17. What is your level of understanding of quantitative methods currently used?

18. What are the main challenges you face under the new policies?

19. What are the approaches that have been working best for you?

20. What do you think the future of anonymisation looks like?