Rare Disease Clinical Data Sharing

phuse.global

Contents

Revision History

Version	Date	Summary
1.0	20 August 2025	Initial Version

Doc ID: WP-095 Working Group: Rare Disease/Small Population Data Sharing Date: 20 August 2025

1: Overview: Purpose of This Document

The purpose of this white paper is to review potential barriers to the sharing of rare disease data such as risk of re-identification and invasion of privacy (balanced against maintaining data utility) to understand if and how these barriers apply to controlled access data sharing under specific contextual assumptions. The development of strategies to enable rare disease data to be shared effectively and be reused is needed to advance research and clinical trial design where there is significant unmet need.

2: Scope

This white paper is divided into three major sections. The first section describes the relevant background to quantitative anonymisation and controlled access platforms. The second section presents features of rare diseases that could impact on re-identification risk, setting the scene for considerations about sensitivity of rare diseases and risk management. The final section discusses rare disease-specific considerations and recommendations. The focus of the white paper is data sharing in a controlled platform context. However, public disclosure of clinical trial data and policies driving this process are used as reference.

3: Problem Statement

Regulatory transparency policies such as EMA Policy 0070 (Phase 1) and Health Canada PRCI have enabled access to rare disease data in a public context, through the availability of clinical study reports and related clinical summary documents. Whilst recognising the value of the progress made in allowing public access to trial methodology, analyses and conclusions there are some limitations to the utility of this data, as except for narratives, individual participant data (IPD) are out of scope.

The biopharmaceutical industry's commitment to transparency and data sharing is reflected in the joint EFPIA–PhRMA Principles for Responsible Clinical Trial Data Sharing released in 2013. Under these principles, there has been an industry-wide commitment to voluntarily share clinical trial datasets, thus making IPD available to qualified researchers for further analysis of controlled data access platforms. However, rare disease clinical trial data is frequently included in trial sponsors' exceptions for sharing statements on the commonly used controlled access platforms Vivli[1] and CSDR[2] and falls outside of any standard company policy.

4: Background

What is a 'rare disease'?

The definition of rare disease varies across jurisdictions, and there is no single definition that is accepted globally. The EMA and Health Canada have adopted the same definition for rare disease: one that affects fewer than 5 in 10,000 people in the European Union/Canada (i.e. one in 2,000).[3,4,5] In the Orphan Drug Act, the FDA defines rare disease as a condition that affects less than 200,000 people in the United States (translating to a prevalence of less than 8.6 per 10,000 at that time).[6]

A maintained registry of rare disease with prevalence, incidence or reported number of published cases can be found on *Orphanet*.

Term 'rare disease' covers a broad prevalence spectrum

There is considerable variation in the prevalence of diseases that are categorised as 'rare' by these definitions.[7] There are rare disorders that are relatively common and close to the rare disease prevalence threshold, such as Sjögren's syndrome and systemic lupus erythematosus with prevalence of >40 per 100,000. Haemophilia A, Fabry disease, Duchenne muscular dystrophy and idiopathic pulmonary fibrosis are rarer (~7 to 16 per 100,000) but are well studied in clinical trials, which is relevant to the determination of a reference population (see below). However, there are many diseases with a much lower prevalence. A review of health technology assessment processes with a modified approach for ultra-rare diseases shows that agencies have defined ultra-rare diseases as those with a prevalence threshold ranging from 1 in 50,000 to 3 in 100,000.[8] Clearly, a one-size-fits-all approach to the sharing (or otherwise) of rare disease data based on an arbitrary prevalence threshold is not appropriate, and, in some cases, company standards and processes applied to non-rare disease data sharing could be considered.[9]

One feature of rare disease data that also varies across diseases is the perceived sensitivity of the data associated with phenotypic manifestations and associated stigma. However, as with the prevalence rate, rare diseases are characterised by a wide diversity of symptoms. Although there are rare diseases with strong phenotypic manifestations for which visual assessment is commonly used for clinical diagnosis (e.g. neurofibromatosis type 1 (NF1) with facial dysmorphology), there are many rare diseases without physical manifestations associated with stigma, and others (e.g. acute myeloid leukaemia) for which physical manifestations are also associated with a wide range of other diseases.[10] Vice versa, there are many non-rare diseases where the data would be considered particularly sensitive, e.g. HIV.

Since rare diseases should be recognised as presenting a spectrum of risk, there is no one-size-fits-all data sharing procedure for rare disease data to endorse. Instead, we will consider if any clinical trial and disease features raise patient re-identification risk and explore generally applicable data sharing practices that may be used to increase data security, and subsequently data utility. Before thinking about strategies for ensuring patient privacy, it is important to describe the context of the data release we are considering, and how re-identifiability risk is calculated.

Controlled access

The context of a given data access model (where and how the data is being stored, shared and used, and by whom for what purpose) has a direct impact on the risk of re-identification and therefore must be carefully considered when sharing data. Privacy, security and contractual controls on data use and sharing point to a lower overall risk of re-identification, with multiple layers of controls bringing multiplicative benefits to mitigating risk.[11]

Controlled data access is the preferred option used by biopharmaceutical companies to fulfil their commitments to transparency and IPD availability. Controlled data release shares clinical information in a non-public manner, i.e. the data is shared with the intention of only being accessible to certain individuals or organisations. The data is analysed on the data access platform with tools provided, and IPD cannot be downloaded. This type of data access model is encountered in multiple different contexts: broad data sharing platforms, such as Vivli, and non-profit biomedical foundations which are focused on specific rare diseases, such as CHDI (Cure Huntington's Disease Initiative)'s[12] Enroll clinical research platform, as well as the direct sharing of clinical data with researchers in pharmaceutical or healthcare companies.

There are many ways of implementing privacy controls:

- Enforceable data use agreements
- Physical controls (e.g. at the data centre storing the data)
- Named access-based controls, limiting the users with data access
- Limitations on downloading, uploading and linking data
- Requiring research proposals from accredited researchers (usually reviewed and approved by an independent scientific review board) outlining a clear purpose
- Anonymisation
- Minimisation (only sharing the data specifically required)
- Defined data retention period

A key control for data sharing is a contract (a data use agreement, or DUA) between the data requestor and the data provider. Before DUA finalisation and sign- off, the data provider and/or platform provider scientific review board typically perform a review of both the research proposal and the lead researcher profile (professional qualifications, scientific accomplishments). This due diligence process ensures the data requested supports the statistical analysis plan's objectives and verifies that the applicant is a legitimate and appropriately qualified researcher (e.g. for Vivli, the lead researcher is affiliated with an institution and has a statistician on the team). DUAs should include wording to the effect that the requestor cannot legally attempt to re-identify an individual or copy the shared data and move it outside of the platform. Moreover, DUAs prohibit the requestor from sharing data with third parties.

The data provider makes changes to the data itself to reduce the risk of re-identification, i.e. the data is anonymised by removing or transforming identifiers that could be used in combination with other information about a named individual. The next section will detail the quantitative approach to the evaluation of re-identification risk using anonymisation. Even though the paper focuses on risk considerations in controlled access environment scenarios, it is important to highlight that IPD anonymisation is a required step in the public disclosure of clinical study documents such as clinical study reports, driven by EMA Policy 0070[13] and Health Canada PRCI[14], and these regulatory documents will be frequently referenced in this paper. In the external guidance for the implementation of EMA Policy 0070 and Health Canada PRCI, the approach to anonymisation is not prescriptive (both qualitative and quantitative approaches are described). In the context of rare disease, it is acknowledged that there may be 'complexity involved in the anonymization of clinical reports in the case of rare diseases, due to the very low number of trial participants and of overall population' and that 'clinical trials conducted on rare diseases and on small populations may have a high risk of reidentification'. In general, in addition to low prevalence, rare disease patients are considered not only to be at higher risk of re-identification for other reasons, such as the sharing of personal information on social media, but also at higher risk of harm were a re-identification to occur due to stigma associated with physical manifestations and familial impact for inherited disease.[15]

Quantifying the risk of patient re-identification

To calculate risk of re-identification, we must consider two factors:

1) The probability of a re-identification attempt
2) The probability of successful re-identification if an attempt were to occur

Pr(re-id) = Pr(attempt) x Pr(re-id | attempt)

Risk of attempt at re-identification

The first element needed to calculate the overall risk of re-identification is the risk of a re-identification attempt (Pr(attempt)). This relates to the context of the data shared, e.g. who the recipient is, contracts, security and access controls. The second element, the data risk, is based on the data itself, e.g. the variables in the dataset(s). In practice, quantitative models of the risk of re-identification, such as k-anonymity, assume the probability of an attempt is '1' for public release and the overall risk of re-identification is controlled by transforming the data and applying data rules. The threshold selected for the re-identification risk should consider the harm that could occur given the sensitivity of the information in the data.

For controlled access data sharing, three types of re-identification attempts (or attacks) can be considered: a deliberate re-identification attempt, an inadvertent re-identification, and a data breach. A data breach occurs when there is unauthorised access to data, e.g. data is lost or stolen, or security systems are breached by hacking, or there are not adequate controls on the data to prevent unauthorised access. The risk of a data breach on a controlled platform is managed by strict security measures, access controls and restrictions on what data can be downloaded from the platform and onward sharing. With this level of security in place, the likelihood of data breach for controlled access is considered low in the context of other attack types. A deliberate attempt to re-identify a trial participant on a controlled access platform would mean that the data requestor would need to break the terms of the DUA, which assures the requestor cannot legally attempt to re-identify an individual. The probability of a deliberate attempt is also dependent on the means and motivation of the potential adversary (attacker). A trusted qualified researcher with a reputation in a research field would be assumed to have a low probability of attempt. The probability of a deliberate attempt based on a range of motivation and contextual controls was proposed (e.g. 0.05 for low motivation and high-mitigating controls and 0.3 for both motivation and controls assessed as medium). In a non-public context for risk assessments, the probability of attempt should be accounted for to maximise data utility.[16]

The final type of attempt to consider is inadvertent re-identification. This would occur when a researcher working with the data recognises that the information relates to someone they know. The calculation method for probability of a spontaneous recognition considers a value called Dunbar's number.[17] Dunbar's number represents the average number of acquaintances (150) someone has, and an assumption is made that the attacker has background information on this acquaintance which includes quasi identifiers in the dataset.

In the clinical trial context, we can calculate the probability of a random person accessing clinical trial datasets who has an acquaintance with the specific disease under study.

If P represents prevalence (cases/population), then 1-P is the proportion of the population in the region who do not have the disease. If a random person has 150 acquaintances in that region, then the probability of none of these acquaintances having the disease is $(1\text{-}P)^{150}$. Therefore, the probability of at least one of the acquaintances having the disease is $1\text{-}\ (1\text{-}P)^{150}$. With rarer diseases with a lower prevalence, the likelihood of knowing someone with that specific disease is lower.
The overall risk of a single attempt will then be the largest value among these three types:

Pr(attempt)=max(Pr(deliberate attempt), Pr(acquaintance), Pr(breach))

Risk of a successful re-identification if an attempt were to occur

The second component is the risk of successful re-identification of an individual were an attack on the data to occur. This means correctly associating a participant's record(s) with a specific named data subject. Variables that can be used in a re-identification attack are called identifiers. The classification of identifiers has been well covered in the literature, and data transformation options can be reviewed in the PHUSE De-identification Standard for CDISC SDTM.[18] In brief, we can distinguish direct identifiers (variables which are unique for an individual) and quasi (or indirect) identifiers (identifiers which could be used in combination to identify an individual). The risk of re-identification for direct identifiers is assumed to be one. Direct identifiers are typically pseudonymized by recoding (e.g. unique subject IDs which are commonly encountered in structured clinical data) or dropped if they have no utility. Quasi identifiers commonly found in clinical trial data include demographic data, e.g. age, sex and race.

For both qualitative and quantitative anonymisation techniques, data is modified so that the risk of re-identification of a trial participant is minimised. Techniques include dropping quasi identifiers with no utility, transformation such as aggregation, offsetting dates, and suppression of individual records.[19] Qualitative determinations of whether a dataset is anonymised are made based on assessments of the transformed data and release/use context provided by subject matter experts. A qualitative risk assessment corresponds to a subjective assessment of the probability of re-identification, taking into account aspects of the subject population and study. This probability tends to be mapped to a scale of low, medium or high risk, which guides the level of anonymisation to apply.

Quantitative determinations generally rely on computing a metric (or metrics) measuring the risk of re-identification of data subjects using mathematical or statistical models. The computed re-identification risk is compared to a pre-determined threshold to determine whether it is suitably low to facilitate sharing of the transformed data. Participants who have the same values for a pre-defined set of quasi identifiers are called an equivalence class. The risk of re-identification of a patient in a dataset can be measured as 1 divided by the size of the equivalence class. This can be interpreted as the probability of selecting the correct record by chance from an equivalence class for a named individual known (because of their combination of quasi identifiers) to be within that equivalence class. A commonly used quantitative anonymisation approach is k-anonymity. In this model, variables are either suppressed or generalised until there are at least k individuals in the dataset which are identical based on the set of quasi identifiers (k being the number of participants in the equivalence class) for all equivalence classes. The probability of re-identification in a k-anonymised data equals 1/k, and this value is called the risk threshold. Larger equivalence classes, and therefore smaller risk thresholds, are associated with a lower risk of re-identification.[20]

Under EMA Policy 0070, for rare disease, a quantitative approach is preferred: "…. a thorough risk assessment should be performed in these cases and anonymization of the data should be adapted to the identified risk." For public data releases, both EMA Policy 0070 and Health Canada PRCI have established a re-identification risk threshold of 0.09 as a baseline in the public context (where the probability of attack is assumed to be 1), which can be interpreted as a minimum equivalence class size of 11 for each data subject.[13] Currently, a common approach of clinical trial sponsors is to limit identifiers to demographic data in the quantitative risk assessments of submissions published on EMA and Health Canada portals, yet it is recognised that low-frequency events, e.g. adverse events and other variables such as dates, must also be taken into consideration when anonymising clinical datasets. In this case, rules-based transformations are applied to the data in addition to the quantitative model.

Quantitative risk models can model the risk of re-identification using just the dataset itself. The underlying assumption in this scenario is that of prosecutor risk, i.e. an attacker would know that a target is in the dataset. One of the situations in which this is applicable is when the dataset is not a sample but rather represents the whole population. In practice, in the clinical trial context, the prosecutor risk is selected when the data provider acknowledges that detailed study population characteristics as per explicit inclusion criteria, as well as narrow timeframe and/or geographical location, make study population distinct from a larger population.

Alternatively, when data providers deem the potential adversary to not know if a participant was recruited into a specific study – for example, characteristics of the study do not differentiate it from a larger disease population – a journalist risk model would usually apply. Under the journalist risk model, quantitative assessments should take into consideration a reference population, which represents the larger population from which the trial dataset is a sample. When the dataset is assumed to be a sample, the equivalence classes are estimated from the whole population using the sample data for estimates of frequencies of and correlation between quasi identifiers. This means that, in general, with a reference population greater than study size, more utility can be retained in the data compared to when the trial dataset is treated as the whole population. The sponsor's assessment of an appropriate reference population for anonymisation of documents for public sharing can differ across companies for similar studies. Options for the choice of reference population for clinical trial data include:

- Study size – the most conservative option, this assumes an adversary knows a given data subject participated in a specific clinical trial, or where the trial population is considered the whole population (can be considered a 'default' option if no other choices are deemed to be defensible)
- Pooled population of related studies across a clinical development programme in a disease area
- Number of individuals who participated in similar clinical trials conducted over an overlapping time frame and with similar inclusion/exclusion criteria – assumes an adversary knows that a given individual participated in a contemporaneous clinical trial in the same indication.

Once the re-identification risk for each subject has been calculated, the overall risk for the entire dataset can be either considered as the maximum risk (based on the smallest equivalence class) or the average risk across all participants in the dataset. The average risk will always be less than or equal to the maximum risk. When data is disclosed publicly, the maximum risk option is recommended, and this aligns with the concept that an adversary targeting any subject, as in a demonstration attack, would target the participants who are easiest to re-identify, i.e. those in the smallest equivalence class. For controlled access sharing, the less conservative average risk model could be used, where the risk of re-identification is the average value across all patients, assuming that an attack on any participant would be equally likely, e.g. in an inadvertent re-identification. An alternative to average risk is strict average risk, which ensures that as well as meeting the average risk threshold, a lower maximum risk threshold is also met.[21]

Another model, which builds on k-anonymity, is l-diversity. Within the minimum group size k, l-diversity requires at least 'l' different values of sensitive variables ("sensitive" meaning an attribute that adversaries must not be allowed to associate with an individual in the dataset). l-diversity is concerned with attribute disclosure and assuming a particular attribute about a trial participant with high probability, if for example the vast majority/all the participants within an equivalence class have the same value for a variable. It can be challenging to achieve k-anonymity (or l-diversity) in a dataset with distinct outliers while maintaining sufficient data utility. Outliers can be completely removed (record suppression), or quasi identifiers can be removed or generalised for specific individuals; however, this may introduce complete information loss about specific groups, e.g. minority race groups, trial participants aged >80. Handling outliers in the dataset in this way must be balanced against retaining greater data utility in the remaining records by reducing the aggregation needed, for example for a variable across all remaining records, to meet a specific risk threshold.

Rare disease: are there factors that could increase the risk of re-identification on a controlled access platform?

It is assumed in general that rare disease trial participants bear greater risk of re-identification compared to non-rare disease participants.[15,23] The following section provides a review of rare disease-specific factors to consider when assessing re-identification risk and planning data sharing in the controlled access context.

Disease population

Disease population is accounted in the inadvertent (Pr[acquaintance]) computation. This shows that the lower the prevalence of a disease, the lower the likelihood of knowing someone with that specific disease, meaning that probability of an inadvertent re-identification of an acquaintance should be lower for rare diseases. However, the risk of an inadvertent re-identification attempt on a clinical trial dataset needs to consider that researchers accessing the data may be clinicians specialising in the disease area, and one or more of the participants may be their own patients. If a clinician believes a record belongs to their own patient, a reasonable assumption can be made that nothing new will be learnt from a clinical dataset that was not already known to the adversary, and we do not consider this to be a re-identification. In addition, moving from a suspected accidental re-identification to a confirmed re-identification would require the researcher to actively check their background knowledge against the participant's information in the dataset, breaking the terms of the DSA.

There are multiple choices for reference population selection for a quantitative risk model (prosecutor or journalist). None of the approaches for reference population selection is specifically recommended by health authorities nor set as a standard within industry recommendations. Data providers individually assess the appropriateness of the reference population selected based on context for both regulatory disclosures and for transparency initiatives. Frequently, for rare diseases, the data provider decides to use the single study population as the reference, which may be appropriate, e.g. a single-site study in a rare disease, or a novel study in a rare disease subgroup. However, for some rare disease studies, there may be a useful contemporaneous clinical trial reference population with similar eligibility criteria. In this case, the reference population will be impacted by how well studied the disease area is, as well as the size of the original study and those studies in the reference population. Alternatively, prevalence and geographic location could be used to estimate the population from which the study is a sample.

Disease population is an important risk factor that needs to be accounted for in the standard risk calculation. When the reference population selected is other than study size (the most conservative option), the parameters used to define the pool of trial participants from which the study is a sample must be carefully considered. This is especially true for rare disease studies. For a well-studied non-rare disease, modifications (such as removal or addition of one or two studies) to a large reference population will likely have no impact on the utility that can be retained in the data. For a small reference population (as is often the case with rare disease), adding or removing just one or two contemporaneous studies can significantly impact on data utility. Strategies will be discussed in a dedicated recommendation section below.

Single-site studies

A review by this working group of anonymisation reports available on the EMA portal shows that clinical trial sponsors using a qualitative risk assessment methodology tended to categorise studies from low to high risk, with high risk frequently associated with studies conducted at a single site or with small study population (usually <100). Single-site studies at specialist centres are often used in rare disease research, which can reveal location information, hospital and treating physician. Rare disease studies by their nature are a challenge to recruit for. However, with cross-border mobility and remote monitoring, many single centres can support international patient recruitment. Therefore, although single-site studies do not necessarily indicate a physical patient location or enable nationality/race to be inferred, this would need to be carefully assessed.

The EMA external guidance on Policy 0070 highlights the impact of location information on the risk of re-identification, stating that: *"A feature of anonymisation is that it is only partially determined by the data to be protected. The ability to identify a trial participant depends on both these data and the state of knowledge of the observer concerning the trial participants in the data. For these reasons, location and dates are important. They may not be the most specific identifiers of a trial participant but they are often the most easily obtainable from other sources. Therefore, clinical data containing information on geographical location and dates should be carefully anonymized."*[13]

k-anonymity-based quantitative risk evaluation can consider patient location information, e.g. country, and its uniqueness in the population. Trial participants recruited from a single site will generally form a homogenous study population as far as this attribute is considered. Where participants all share an identifier such as country or site, the patient location will have no impact on risk calculations when a prosecutor risk model is selected. In fact, the data provider should consider retaining the identifier since it is knowable from other sources, such as protocol or clinical study report. Considering a scenario where a data provider assumed a journalist risk model – for single-site studies where location down to the specific clinic is public information – the reference population could be based on disease prevalence and population estimates from the recruiting geographic region, which can also be narrowed using eligibility criteria such as age range and gender.

In general, the intersection of rare disease and a single site is a situation in which it may prove difficult to maintain sufficient data utility using only quantitative methods to protect participant privacy, even in the controlled access context.

Social media activity

Previous studies of rare diseases have confirmed the importance of patient organisations and knowledge sharing with other people experiencing the same condition. [24,25,26,27] Apart from dedicated online support forums, it has been recognised that people commonly display their personal information freely on social media. This sharing of personal information occurs in both completely open and closed groups.[28] Social media can be used for understanding newly diagnosed patients' conditions and prognoses, sharing information about new treatments and research developments, finding medical specialists, emotional support, and help with everyday practical health issues. Patients and families sharing information on public or even private online support forums could put them at increased risk of privacy breaches. Even high-level information ("I live in mid-Western US") could contribute to a re-identification attack, especially when considering public disclosure of patient data.

A US national survey in 2011 found that approximately one in four internet users living with a chronic health condition (e.g. high blood pressure, diabetes, heart condition, lung condition, cancer) had gone online to connect with people with the same condition. This national survey suggested that for rare diseases there is an increased social media presence for health-related purposes and a reliance on extended networks, with over half contacting others with the same health condition. However, the authors acknowledge that for rare diseases the population targeted was composed of people who were already part of a rare disease online community, meaning the results could not be directly compared with chronic disease in general. Moreover, the use of social media for health-related purposes is not specific to rare diseases and is widespread.[29,30] For example, there is evidence that the COVID pandemic impacted on the sharing of health information, with personal information shared more readily, and sharing not only COVID-19 diagnoses but preexisting health conditions that made them potentially more vulnerable.[31,32] However, it must be acknowledged that rare disease patients are less likely to have access to people in support groups, for example, who live near to them and share their condition, which could impact on their social media presence, especially around their condition and treatment.

Given that social media presence for health conditions is widespread, the question arises of whether there is a greater impact on privacy risk for rare diseases in a controlled access setting relative to other conditions. Rare diseases cannot be treated as one bucket of greater social media use. The increased re-identification risk from social media presence will vary based on disease, study location and trial participant characteristics. In a controlled access environment, a contracted agreement between data contributor and requestor prohibits intentional patient re-identification. Irrespective of the disease setting and its prevalence, malicious use of information left privately on the internet by a patient in combination with clinical data obtained via a trusted data sharing model in a re-identification attempt is highly unlikely. Hence, this paper's authors assess the risk of a researcher using social media information to be of low risk, and not specific to a rare disease setting.

Other information in the public domain: publications and transparency regulations

A scientist coming up with a research idea and then writing a research proposal will inevitably have explored data of interest already available in a public domain (e.g. a full study report if available on health agency portals). Moreover, clinical study reports are frequently requested together with clinical trial datasets and delivered as a part of data packages. Harmonisation practices supporting consistency across transparency initiatives have recently become a subject of debate. For example, safety information on an individual trial participant can be found in CSRs, summaries of safety, safety update reports, PBRERs (periodic benefit-risk evaluation reports), publications including case studies, and registries; meaning more information about an individual may be derived when information is looked at in combination, especially where there are inconsistencies in approach.

One example for difficult-to-recruit rare disease studies occurs when there is a need to set up multiple sites in different countries, which can result in small numbers of trial participants at each location. Slow recruitment and a low number of patients at each site can impact on re-identification risk when multiple snapshots of the data are taken for interim analysis or when recruitment is kept open after primary analysis until a threshold number of patients within a subgroup of specific interest is achieved. If one or a small number of additional participants are recruited between snapshots, it may be possible to create individual profiles from summary tables in clinical study reports shared alongside the data (or publicly available) including demographics, recruiting physicians (and therefore location information more granular than country), adverse events, and medical history, which will impact on the risk mitigation steps taken through anonymisation of the datasets.

A comparison of summary tables in CSRs and in registries based on a difference in population 'N' can also allow profiles to be built for participants who drop in and out of populations, e.g. 'All Patients' in a registry versus a 'Safety Population' in a CSR. Another example is a detailed case study style publication of a trial participant, including safety information, with contributing physician and affiliation, which, when compared to a narrative, can render redaction of location information in the narrative ineffective.

A set of Kaplan Meier plots by multiple demographic subgroups could allow race, sex and age group to be linked for specific participants, particularly if the data is sparse and there are unique event times. Similarly, a box plot repeated by demographic subgroups could allow information to be linked to outliers.

While none of these examples is specific to rare disease, rare disease studies may be more at risk of triangulation of information due to their small size, slow recruitment and a heightened interest in specific individuals for case studies. This means that due diligence in assessing the full breadth of information available on the study to the public (as well as planned information sharing such as upcoming publications) is an important part of the risk assessment when sharing rare disease clinical trial datasets. A quantitative risk assessment must be valid in the context of the totality of the information available to the researcher accessing the data.

Sensitive information and identifying disease traits

Clinical datasets are often complex and contain a high volume of detailed clinical information, including reported adverse events, concomitant medication and patient medical history. Therefore, anonymisation pertains to multidimensional data transformation and is not limited to the recoding of subject identifiers and the handling of quasi identifiers within demographic data.

Adverse events, medical history records (and related concomitant medication) and detailed characteristics presented as investigator text can serve to single out a trial participant and thus may be considered quasi identifiers dependent on the study context, e.g. the disease area. These data types are sometimes described as identifying and/or sensitive, and this terminology is often used interchangeably without clarity, especially for adverse events. Here, we use the terminology 'sensitive' to mean that if a re-identification were to occur and the sensitive information were disclosed, the negative impact on the trial participant could be significant (e.g. embarrassment and impact on wellbeing in terms of employability, reputation, insurability, self-esteem, stigma, or loss of income).[33] Examples include substance abuse, mental disorders and abortion. A sensitive event may not be commonly known or visually identifying, e.g. a sexually transmitted disease such as HIV or gonorrhoea. Unusual and visually identifying adverse events can also be sensitive, e.g. body deformations (amputations, kyphosis), but that is not always the case. Once an event has been assessed as sensitive and/or unusual/visually identifying, the coded term should be considered to potentially require protection prior to data sharing. In general, verbatim text and investigator terms should be removed (as they are not usually required from a utility perspective and may contain specific or unusual anecdotal details), and only the corresponding coded terms retained and reviewed for sensitivity.

There are different de-identification techniques used by data providers to manage sensitive and/or potentially identifying events in addition to quantitative risk modelling. One potential methodology is based on subjectively created lists of such sensitive events and finding the overlap in the data, i.e. a rules-based approach. Information is either blanked or replaced by the text, indicating data removal. Another, not mutually exclusive, commonly used technique is assessing terms with low frequencies in a population with subsequent removal of identified rare events that could contribute to the singling out of a trial participant.[18]

Removal of all cases of a sensitive/identifying event irrespective of its distribution in study population is not always appropriate. One suggestion is that sensitive/identifying term lists should vary by disease area.[34] This would be appropriate, for example, if the disease was associated with the term, so by inference there is a high probability that the sensitive term would apply to any participant and could be removed from the term list, e.g. if mastectomy identified as sensitive/identifying was included in a default term list, all mastectomy cases in a breast cancer study would be redacted, even if the study population was found to be homogenous in terms of this attribute distribution. Modifying the term lists based on study frequency is especially valid when a prosecutor risk model is applied. On the other hand, sensitive/identifying attributes shared by study participants, e.g. based on eligibility criteria, may distinguish the study population from the general disease population and impact on whether alternative journalist risk model selection is appropriate. The re-identification potential of an event or medical history should also be considered in the context of commonality in the disease population. Removal of terms based on frequency alone within a study can result in significant data removal from an AE dataset that is not necessary, considering a dataset may include many low-frequency events reported in the study population, but which are common in the general population, e.g. diarrhoea or fever. To conclude, automated solutions will negatively impact to some extent on data utility. A hybrid solution of using a term list adjusted to the disease area and using objective criteria based on event frequency is recommended. For example, event terms with frequency above a specific threshold could be retained regardless of sensitivity, whereas those falling below the threshold would be reviewed for sensitivity against a disease-specific rule set.[35] For low-frequency events, consider the reference population chosen. If using a prosecutor model, low-frequency events/unique events would be determined from within the dataset itself, or, if using studies within a clinical development programme, the overall safety database could be used to estimate the frequency of events in the disease population.

Fully quantitative approaches to adverse event data are possible (e.g. using a separate risk model for the AE data and aggregating preferred terms to higher-level terms) but come at the expense of retaining any real utility, since the combination of events for any one participant is often unique. Some data providers are implementing l-diversity for sensitive attributes to assess whether variability is sufficient within the equivalence classes; otherwise, the attribute value is removed.

The number of potential approaches for medical event anonymisation is indicative of how difficult it is to find a balance between automation and time-efficient solutions supporting fast-track anonymised data provision versus maintaining high data utility and appropriate handling of re-identification risks.

Some rare disease datasets are highly sensitive. Rare disease stigma occurs when a certain attribute or identity is deemed socially unacceptable or inferior, leading to structural and interpersonal discrimination.[36] This discrimination contributes to social inequity and can negatively impact on those with the stigmatised trait. Previous research studies have noted that stigma can be a common issue for people with rare diseases.[37] It should be noted that rare disease patients often suffer from social and psychological challenges. Their conditions may serve as an object of unhealthy fascination and discussion in their community, e.g. visible attributes of rare diseases can be taken to be transmittable and attract the curiosity of both acquaintances and strangers. Rare disease stigma can therefore increase the risk of a re-identification attack on its own. Among such disorders are orofacial abnormalities and psoriasis (one of the most common rare diseases, often correlated with visible inflammation on the hands). Rare disease patient records may be more sensitive than non-rare disease patient data. Regardless, it is essential that medical history, adverse event and concomitant medication datasets are thoroughly analysed and processed accordingly.

Genomic data

It is estimated that 72% of rare diseases have a genetic cause, while others are the result of infections, allergies and environment.[5] Clinical trial datasets in rare diseases frequently contain genetic information. In clinical research into rare diseases, genetic testing may be used to identify genetic variants associated with the disease under study, and to further understand how they may be associated with the natural history of the disease and the response to treatment. Genetic information may be used to select trial participants for inclusion in studies, to ensure balance of disease subtypes, or for exploratory subgroup analysis. Genomic data may be considered higher risk for sharing than other types of clinical data; some DNA-based lab test results can be replicable for an individual over their lifespan. Information about an individual's current health/risk of future health issues, behaviour and phenotype (appearance) can be inferred from genetic data and therefore be highly sensitive.

The privacy risks associated with the sharing of genomic data vary based on the type and extent of the genetic data being shared. The uniqueness of some types of genetic data does not mean that all genetic data is inherently re-identifiable. As with other types of data, the re-identifiability of data depends upon being able to associate it with a reference dataset relating to the same individual that contains additional quasi and/or direct identifiers. Importantly, not all types of genetic data uniquely identify an individual or are replicable, hence there is no uniform re-identification risk level associated with genetic data sharing that would be applicable to genetic data in general. For example, gene expression count data would be considered less reproducible/static than sequencing data or variant data and does not contain direct information about genotypes. The replicability of genetic data is also dependent on the sample used to derive the genetic data. Genetic data derived from tumour cells will contain both inherited variants and acquired somatic information (e.g. mutations), and tumour samples can be a mixture of cell types, which evolve over time and acquire more mutation and introduce variability within an individual. A review of privacy attacks on genetic data identified nine features that inform privacy risk of genetic data: biological modality, experimental assay, data format or level of processing, germline versus somatic variation content, content of single nucleotide polymorphisms, short tandem repeats, aggregated sample measures, structural variants, and rare single nucleotide variants.[38]

The inherited genetic component of many rare diseases and the potential for sensitive information to be inferred from genetic data about blood relatives means the privacy risk extends beyond the trial participant and brings increased risk to data sharing in terms of identifiability and potential harm. In some cases, trial participants and family members may not know or want to know their genetic status and predisposition to an inherited disease that has not yet manifested. However, in rare diseases, single-gene causality or diseases characterised by a low complexity of genetic pattern are relatively uncommon. Cystic fibrosis, sickle cell anaemia and Tay-Sachs disease are examples of autosomal recessive rare disorders that happen in the offspring of couples who are both carriers of potentially lethal gene variants. Haemophilia, Duchenne muscular dystrophy and X-linked mental retardation are passed from maternal carriers and affect hemizygous males.

Kyoko Takashima et al. deliberate on familial disease data sharing when the patients' family members data is also collected and shared.[39] The authors present survey results conducted among healthy Japanese adults and patients, showing public expectation that familial data will fall under stricter data privacy protection procedures. Moreover, they share general recommendations for safeguarding privacy for familial diseases. One of their recommendations for data stewardship is careful construction of informed consent, diligence in explaining the secondary use concept, and presenting measures to be undertaken to protect their privacy, including privacy of relatives.[39]

In January 2018, the International Rare Diseases Research Consortium (IRDiRC) and the Global Alliance for Genomics and Health (GA4GH) developed model consent clauses tailored to rare disease research, including important clauses for rare disease research settings and complementing classic consent forms. Consortia emphasise that "the challenge in establishing consent policies for rare disease research stems from the dichotomy between the push for free flow of data against concerns about loss of privacy." The key implications are inclusion of family data, complexity of collected data driven by development in research technologies, e.g. genetic and phenotypic data including audiovisual data collection. The complexity of data collected means that consent forms may become overcomplicated, and care must be taken to make clear the study purpose and possible benefits, as well as identifiability risks (in the context of sociodemographic, family history, genetic and phenotypic data) and disclosing privacy protections for the use and sharing of family data.

5: Recommendation

To lower the risk of re-identification to an acceptable level, transformation of the data resulting from standard quantitative methodology practices may lead to low utility, which is a barrier to data sharing. This can be perceived as inevitable for rare diseases. However, this will not always be the case, especially for diseases that are not ultra-rare. The data release context can significantly impact on data utility by allowing a less conservative choice of risk threshold compared to public disclosure practices. Model assumptions made for non-rare disease shared in the same context should not by default be made more restrictive, e.g. automatically choosing a more conservative risk threshold when the disease is rare and treating the study population as the total population as a default. Reducing the sensitivity of the data and retaining only data that is required for analysis by data minimisation, e.g. dropping datasets/variables and removing sensitive terms, may allow a less conservative risk threshold to be used and/or reduce the number of quasi identifiers in the dataset.

Due diligence process for managing data sharing

Controlled access platforms manage risks by embracing multiple controls. However, it is good practice for company transparency teams to carefully manage a due diligence process to ensure their requirements are met across platforms. Frequently, contracts are only signed by the lead researcher, whereas extra controls in rare disease data sharing may be warranted, such as contractual agreement by all researchers accessing the data and working on a publication. Many controlled access platforms allow the downloading of summary data for use in publications. Courtesy review of publications for scientific content could include a privacy review, as the data will be shared publicly, to check whether the data was anonymised for the controlled access context (e.g. check for individual participant data in profile plots, review cross-tabulations by demographics). The data provider could review this information as an extra step for particularly sensitive rare disease datasets in the context of other publicly available data, such as transparency registries and sharing of documents under regulatory requirements, such as EMA Policy 0070 and Health Canada PRCI, because different anonymisation approaches, rather than a holistic one, for IPD in documents may increase the overall re-identification risk.

Platforms such as Vivli give the data provider the opportunity to connect with the data requestor. Upon review of the research proposal, questions can be asked of the researcher. Not all sharing requests are approved, and decisions can be made on a case-by-case basis for rare diseases. For example, a request can be rejected if qualifications and affiliation are not aligned with expectations based on the proposed research plan. For rare disease studies, more interaction between the data provider and the researcher may be required, because a thorough understanding of the research plan will be needed to provide tailored data with potential data minimisation techniques, including only the data types required for analysis (see below).

The anonymisation of datasets needs to align with previous and planned data releases by the provider in documents, registries and publications. A review of released information should inform the anonymisation approach, or, ideally, the approach to sharing the dataset should be considered prior to anonymising documents. Once data has been released, the granularity of a specific variable in the public domain dictates future releases, because subsequent transformations can be more, but not less, conservative. Typically, sharing IPD as part of regulatory disclosure would occur prior to making the datasets available under transparency initiatives. Often in patient narratives there is a very limited set of quasi identifiers, e.g. race, sex, age, sometimes country, body weight or BMI. Under EMA Policy 0070 and Health Canada PRCI, priority information to retain (dependent on context) in regulatory submission is generally the adverse event data, medical history and concomitant medication. When working with the data with a quantitative model to define, the transformations that will be applied to documents, if only the quasi identifiers within IPD in the documents are considered, can be problematic when it comes to subsequently sharing the underlying datasets and prioritising information retention. For example, in a quantitative model containing age, sex and race only as quasi identifiers, race may be fully retained. For the same model assumptions and threshold, if more quasi identifiers are included (those required in the dataset, e.g. country, but not present in the narratives), then race may need to be aggregated in the optimal model. However, limitations on model selection as 'race' without transformation have already been released. In addition, the order of prioritisation of quasi identifiers may vary between a clinical study report and a bespoke anonymisation of a dataset dependent on analyses in the clinical study report and a specific research proposal. These types of misalignments are more acute in datasets where greater transformation and careful balancing of risk across the variables is required, such as the rare disease setting. When transforming data for the subsequent anonymisation of documents, other important quasi identifiers specific to the disease area, not present in the documents in IPD, could be factored into the model. Using a less conservative risk threshold for controlled data access as opposed to public release will also help minimise these issues. (See the risk threshold section below.)

Well-selected model and reference population

Data anonymisation standards available are not prescriptive on what is the rightful reference population for a risk assessment. The recently published TransCelerate resource on Privacy Methodology for Data Sharing recommends demography data be removed based on frequency measurement (data below a frequency threshold is redacted), but there are no specifics to the calculation itself mentioned, i.e. is this only applicable in the case where there is no reference population other than the trial population itself?[35] This may be interpreted as a reflection of the current state of industry practices, with such a range of reference population types used for risk assessments and accepted by regulators. The authors of this paper will not attempt to convince the superiority of one approach over the other, rather reflect if any changes to standard risk assessment calculation should be considered when managing rare disease data.

Selecting the single-study population as the reference is the most straightforward, least resource-intensive, safest and most conservative option, but it may affect data utility to a level of not being of any use to the research proposal, with the highest probability of data below a low-frequency threshold, especially for smaller studies. If the data contributor has an established approach to individuals who participated in similar clinical trials used as the reference, then this approach could also be followed for rare disease. Nonetheless, in the rare disease setting, careful documentation of the parameters that feed into the reference population selection should be made before selecting similar trials to avoid adjustments based on the outcome, e.g. broadening the disease area definition to be more inclusive if no similar trials are found.

Typically, when a journalist risk model is selected, trials are selected with matching eligibility criteria, overlapping timeframe and geographic location for participant recruitment, and this can be semi-automated from clinical trial registries. For rare disease studies, even though more laborious, a manual review of reference studies could be performed, and this might be manageable from a resource perspective due to the small number of studies to be considered. Care must be taken to ensure patients are not represented more than once in the reference population, for example in an extension study under a separate protocol, which could be picked up with a manual review. Manual review of similar trial data could identify studies with limited overlap in demography or disease characteristic, e.g. a study could be removed from the reference that was eligible for all-comers, but where the vast majority had mild to moderate disease if the study to be anonymised had recruited mainly severe patients. Subsets of participants could be included in the reference 'N' rather than the whole study, e.g. an arm from a basket study.

The pooling of studies in a development programme is recommended when there is overlap in eligibility criteria, even if only one study is to be shared. This not only supports a combined reference population, but model estimates of frequency and correlation of quasi identifiers derived from a larger population with the quantitative model may also contribute to better data retention and could be impactful in the rare disease setting.

With a journalist risk model selected as per a data provider's standard approach, in a rare disease setting and when participants are localised in a single country or site, the reference population should take this into account. Including reference studies with limited overlap in location is not recommended. For rare disease studies run out from a specialist centre, it may be possible to determine a reference population based on concurrent clinical trials run out of the centre in a specific disease area. This information can be found by filtering for centre name on clinical trial registries such as ClinicalTrials.gov. Alternatively, many sites have their own public-facing website listing ongoing clinical research programmes and trials. A prosecutor model could also be considered.

Handling outliers – multimodal model

When there are small cell sizes and a problematic distribution of participant attributes (resulting in unique trial participants for a combination of quasi identifiers, referred to as outliers), a small number of records can drive information loss when using a quantitative approach to anonymising a dataset. In such a scenario, the grouping of participants into the required equivalence classes may be unfeasible without a more conservative approach, e.g. generalisation to higher order, wider banding. One option is participant suppression, in which all the quasi identifiers for the outliers are removed or the records dropped completely. For the former, it is important to ensure a comparison of the dataset to the anonymised or redacted documents does not negate the removal of the information in the dataset or allow inferences to be drawn about the suppressed participant. For example, if sex was a variable with a low-frequency group, such as male patients in a breast cancer study, it would be reasonable to assume that the male patients were among the participants with redacted identifiers, even if other outliers were included in this outlier subset. Alternatively, to avoid insufficient data retention, a multimodal k-anonymity risk analysis could be applied. In a multimodal model, groups of participants are handled differently, with sets of anonymisation methods applied simultaneously, and k-anonymity is achieved within each set. This type of model can achieve greater data utility for a major subpopulation in the dataset, at the same time as sacrificing data utility in minor subpopulation(s) that involves problematic outliers. In this case, instead of applying the same transformation for a variable across all participants, subsets of trial participants may have a different transformation applied, e.g. for most participants, five-year age banding may be used, whereas 20-year age banding may be used for a minority subset. Preserving more data utility in trial participants at the expense of a smaller subset may be a useful strategy in rare disease dataset anonymisation. However, this depends on the research project's focus, because it may be preferable to maximise information on outliers for some variables, with more information loss across the study population for others. Again, an in-depth understanding of the research proposal is needed when making these decisions about model options.

Data minimisation

Following a company-standard approach may not be possible for successful rare disease data anonymisation. A deep understanding of the disease area, the data and any identifiers, and the specific research proposal will likely be needed to successfully anonymise the data, while ensuring it is fit for purpose. Bespoke solutions will likely be required for each research proposal. The study of rare diseases often involves deducting correlative or causative genetic relationships to disease traits and progression, and a high degree of data utility may be required for certain variables which are quasi identifiers. For example, the severity of Huntington's disease and the age of onset have a relationship to CAG repeat length (a genetic sequence repeated more times in people with Huntington's disease than without).[40] Therefore, more background knowledge of the disease from transparency experts and an upfront effort to understand such data utility requirements (i.e. in this example, prioritising retaining age and CAG repeat length) are important when anonymising the data. Identifiers which would normally be transformed may be a priority to retain. Because of this, companies may choose to not proactively list rare disease datasets but anonymise and share the data only with a specific request in mind.

The clinical data collected may be very detailed and abundant in sensitive information. If information is not critical to understanding or interpreting the study results, it should be considered for removal from the data to be shared so that a higher risk threshold can be used, compared to a situation where such sensitive information was retained. Special care should be taken when unusual events are also of a sensitive nature, with the possibility of causing harm when associated with a named individual. Apart from adverse events, it should be acknowledged that medical history events, even if common in the general population, may present a unique combination of events that could be known by an acquaintance and contribute to the singling out of an individual. A preventive action to minimise the risk is the complete or partial exclusion of some datasets, e.g. medical history (perhaps retaining those terms related to the disease, e.g. coded signs and symptoms), surgical history, and previous medications not concurrent with the study, if not indicated as a priority within a data request form. For an efficacy-related research proposal, none of these data types may be required.

Clinical trials on rare diseases will often recruit participants with more than one rare disorder or with a disorder with multiple subtypes, for some of which prevalence is estimated by number of known cases, or number of family groups, e.g. histiocytic disorders with subtypes Erdheim–Chester, Rosai–Dorfman, xanthogranuloma, or Langerhans histiocytosis. Differentiating identifying/sensitive information of subtype of the disease may be removed depending on the distribution of the subtype and threshold selected, i.e. the subtype may be treated as a quasi identifier.

If genomic data is not required as per a data sharing request for a research purpose, it should be considered for exclusion. Nevertheless, we discourage an automatic rejection of such data sharing requests based on a valid justification and research plan. Instead, a risk assessment should be performed considering minimisation, aggregation and using summarised data types where possible, and removal of quasi identifiers such as demographics where not required. For example, variant data can be presented in an aggregated form so that the proportion of data subjects with a specific variant is shared. Variant information from tumour samples can be expressed as the differential between control and tumour genotypes so that only acquired variation is shared. Transparency experts may not be experts in genomic data types and their potential contribution to identifiability in the context of the disease area. An open dialogue with internal experts on these data types may be needed to understand, for example, the frequency of variants and their combinations in the disease area to make considered decisions.

Data minimisation practices are also still widely used in individual participant-level data anonymisation within clinical document disclosure, e.g. the heavy redaction of case narratives. Regulatory bodies, however, caution against its overuse. Extensive redactions are particularly applied in sensitive settings, hence may be considered upon rare disease clinical report anonymisation. When prioritising the retention of individual participant-level data in the narratives, there may be an increased need for redaction at the summary level to prevent linking information about individuals across data representations, which could compromise risk mitigation. Additionally, a more conservative approach may need to be taken with quasi identifiers within the narrative by choosing a more conservative quantitative model for the demographic information.

Risk threshold selection

As noted previously, both EMA Policy 0070 and Health Canada PRCI have established a default risk threshold of 0.09 for the maximum risk metric, referring to this as a 'conservative threshold' for public release. A review of anonymisation reports for rare disease clinical trials on the EMA Clinical Data Publication portal indicated that sponsor organisations, when employing a quantitative approach, most often chose a threshold of 0.09, in line with the threshold recommended by the EMA.[41] Accordingly, there is some evidence to indicate that regulators accept a quantitative assessment made with a choice of threshold of 0.09 for rare disease data. In the absence of a strong precedent or explicit guidance, for a strictly quantitative approach to assessing risk of re-identification, it may be worth considering stricter thresholds for rare disease data for public data sharing (i.e. larger equivalence class sizes), based on the sensitivity of the data. For example, it has been suggested that a threshold of 0.05 (an equivalence class of 20) may be appropriate for the release of highly sensitive data.[42] This threshold has also been used for ultra-rare disease regulatory disclosure on the Health Canada PRCI portal.

In the public disclosure context, the risk threshold may be directly translated to the smallest equivalence class allowed in the data. However, for controlled access data sharing, where the probability of attack is usually assumed to be <1, such a simplified interpretation can be highly misleading and should therefore be avoided. The risk threshold value determined for use in controlled data sharing, e.g. 0.09, will not correspond to the equivalence class of the inverse of this value in the transformed data if the probability of attack is also considered when meeting this threshold. It is then good practice to transparently specify both the selected risk threshold for acceptable overall risk of patient re-identification and the minimum equivalence class targeted in the data (or risk of re-identification in the data itself).

By way of example, if targeting an equivalence class (EC)=3 with a quantitative model, assuming a probability of attack of 0.3, this would yield an overall risk value of 0.09, which is repeatedly mentioned as the recommended threshold for public disclosures.[16] The lower probability of attack makes it possible to generalise the data into smaller groups of individuals while still meeting the same overall risk threshold. If targeting EC=2 and assuming a probability of attack to be 0.3, this would give a maximum risk value of 0.15.

El Emam et al. introduced a recommendation that deems the presence of uniques and/or a class of two after de-identification as undesirable (minimum equivalence class of 3). Additionally, they introduce the concept of 'strict average risk' as a two-step measure to ensure the absence of unique or double records in the dataset when using average risk metrics.[16] In the PHUSE deliverable, the authors also contemplate using an additional metric – the 'uniqueness threshold' – when using average risk to ensure the absence of unique records in the dataset.[18]

Within a specific context, the risk threshold can be adjusted for the same study based on the level of sensitivity that will remain in the data (e.g. considering the extent of data minimisation). In addition, the risk threshold used for public release, along with the anonymisation methodology, should be factored into decisions made for controlled access sharing. For example, if a risk threshold of 0.09 (equivalence class of 11) was used for public disclosure of documents (e.g. CSR for a rare disease) – where sensitive data had been redacted – when sharing the data on a controlled access platform (again with sensitive data removed), the same overall risk threshold would be appropriate. However, if the probability of attack was assumed to be 0.3, then the threshold targeted in the model would be 0.3 (0.09/0.3), i.e. an equivalence class of approximately 3. If the sensitive information were to be retained (prioritised as per scientific research requirements) in the controlled access environment, then a lower risk threshold would be appropriate, e.g. 0.05, resulting in a targeted equivalence class of approximately 6 when taking into account the probability of attack. Importantly, the selected methods for controlled access should not contradict those applied to documents for public disclosure. Note, even the lowest recommended value of 0.05 for probability of attack assumes there is a 1 in 20 chance of a deliberate attack, which is very conservative. In the context of a controlled access platform, a strict average risk threshold could also be considered.

Depending on rare disease hallmarks (identifying traits associated with the disease, familial information), study design, historic decisions on sharing through transparency initiatives, and data processing methodology, the data provider should decide whether the risk threshold used for standard data sharing (non-rare disease) in the same context needs to be adjusted. If sensitive information is retained because it is essential for utility, then the risk threshold may need to be more conservative. However, increasing the equivalence class size would typically necessitate implementing a more conservative anonymisation methodology, the impact of which can be high on small, rare disease datasets under the prosecutor model or with a small external reference population. Where key demographic data is critical for secondary use of data (e.g. age in Huntington's disease), a more conservative threshold can render the research project untenable. Risk mitigation based on multimodal anonymisation application to handle outliers is recommended and/or sharing a tailored dataset package with maximum utility maintained on the most important stratifier(s) as indicated in the research plan, dropping other quasi identifiers. Reducing the overall sensitivity of the data, for example by dropping variables and datasets, may mean the standard risk threshold used by the data provider is considered appropriate. Alternatively, if the focus of the research is dependent on sensitive information, it may be necessary to drop most of the demographic data based on the quantitative risk assessment.

Additional technical aspects for consideration: dataset processing methods enhancing data utility

The authors of this paper have frequently referenced the data sharing platform Vivli, one of the most recognised clinical trial sponsor controlled access data repositories. Nonetheless, there are many disease-specific controlled access platforms available for rare diseases that pool data from multiple sources. For example, C-Path is an independent party that enables public–private partnerships of regulatory agencies, biopharmaceutical firms, universities, and patient groups in the sharing of scientific data. The C-Path Duchenne database is curated and integrated so that researchers are not able to identify the specific trial from which individual participant observations originate. Though privacy protection practices in repositories offering metadata are not a key consideration of this paper, the authors recognise the potential of this model. It is worth mentioning that AI-driven alternatives to canonical anonymisation are rapidly developing, although currently there are no systematic reviews that investigate the efficacy of how machine learning is used in a rare disease context. Interestingly, a feasibility study was recently conducted where SMPC (secure multi-party computation) was applied within the Collaboration on Rare Diseases project (CORD_MI), which involved German informatic consortia, German clinics and patient associations. SMPC is a cryptographic protocol that distributes a computation across multiple parties, while keeping the data inputs private.[43] Artificial intelligence machine learning approaches are emerging that use privacy-preserving techniques for distributed computations. One interesting approach is federated learning (FL), where both participants keep their raw data on their premises and exchange intermediate model parameters. To summarise, emerging methodologies may be impactful in the rare disease space where standard approaches have limitations.

Conclusion

It has been noted that the systematic treatment of public disclosure of rare disease data with a stricter approach to anonymisation, resulting in more stringent transformations, may create potential disclosure biases, whereby only releases of non-rare disease data may retain any meaningful utility.[44] The potential benefits (which it could be argued are higher for rare diseases with unmet need and limited data availability[45]) of sharing data need to be balanced against the privacy of a low-risk context such as a controlled access platform.

The characteristics of rare disease studies that may increase re-identification risk are not universal across rare diseases, nor are they restricted to rare diseases. Data providers should consider whether their standard risk model assumptions can be applied to rare disease data – considering use of a reference population if justified – and to handling sensitivity by carefully prioritising information to be retained and data minimisation for each request, and whether increased due diligence, for example full assessment of other publicly available information and a privacy review of downloaded summary data, can reduce the privacy risk. It is acknowledged that where there is an intersection of risk factors, or for small studies in an ultra-rare disease, the quantitative approach will not be feasible and other solutions may need to be considered, such as federated access to data. Sharing rare disease data will often be more resource-intensive if a more thorough understanding of the disease area is required by transparency experts and bespoke data packages are provided. Providers may need to prioritise data requests based on a quality and value assessment because resources for voluntary sharing of data are not limitless.

Disclaimer

The opinions expressed in this document are those of the authors and should not be construed to represent the opinions of PHUSE members, respective companies/organisations or regulators' views or policies. The content in this document should not be interpreted as a data standard and/or information required by regulatory authorities.

Project Contact Information

- Helen Spotswood, Karolina Stępniak, Shalini Dwivedi, Stephen McCawille, David Di Valentino, Neha Srivastava
- workinggroups@phuse.global

Acknowledgements

We would like to thank Christopher Battison, who was involved as Co-Lead from the start of the project until 2023 (when he had to step down from the project due to conflicting priorities), and Jean-Marc Ferran for their invaluable support in shaping this paper.

Glossary

Definitions are taken from the PHUSE Terminology Harmonisation in Data Sharing and Disclosure Deliverables (unless otherwise cited).[33]

Term	Definition
Adversary	A data user who intentionally or inadvertently learns or discloses information about a data subject through re-identification or attribution. This user may be motivated by a wish to discredit or otherwise harm the organisation disseminating the data, to gain notoriety or publicity, or to gain profitable knowledge about data subjects. Data adversaries are sometimes referred to as intruders, snoopers or attackers.
Anonymisation	The overall process of protecting the privacy of data subjects, including clinical study participants, and reducing the risk of re-identification by 1) modifying (e.g. suppressing, obscuring, aggregating, altering) identifiable information in structured data and documents 2) assessing and controlling the residual risk of re-identification 3) considering the context of the data release.
Attribute disclosure	Occurs when a sensitive attribute about a participant in the database can be inferred with a sufficiently high probability.[46]
Controlled access	Requires a request for access to the dataset to be approved. Controlled access limits data sharing to researchers with a specific, relevant research question. The restrictions are determined by the data owner. DUAs are often used in controlled access data sharing.[47]
Demonstration attack	A type of re-identification attempt in which adversaries are most likely interested in showing that an attack is possible.[48]
De-identification	A general term for any process of removing the association between a set of identifying data and a data subject present in data or documents. The association between data and subject is removed by modifying (e.g. removing, obscuring, aggregating, altering) identifiable information in structured data and documents.
Direct identifier	Data that can be used to uniquely identify an individual (e.g. study participant ID, social security number, exact address, telephone number, email address, government-assigned identifier) without additional information or cross-linking other information in the public domain.

Equivalence class	Records (i.e. rows in a dataset) that share the same values for variables in a set of quasi identifiers.
Generalisation and data aggregation	Diluting the attributes of data subjects by modifying the respective scale or order of magnitude (i.e. a region rather than a city, a month rather than week).[49] Aggregation techniques aim to prevent a data subject from being singled out by grouping them with other individuals. To achieve this, the attribute values are generalised to such an extent that individuals share the same values.[49]
Individual participant data (IPD)	The person-specific data separately recorded for each data subject in a clinical study.
Journalist risk	The risk of an adversary (individual or organisation) intentionally attempting to identify a data subject within a dataset. The adversary does not know if a specific individual is in the dataset.
k-anonymity	A criterion used to ensure there are at least k records within each equivalence class in a dataset.
l-diversity	A refinement to the k-anonymity approach which assures that groups of records specified by the same identifiers have sufficient diversity to prevent inferential disclosure.
Offsetting	A technique to anonymise dates, in which a random offset is generated and applied to all dates. All original dates are replaced with the new dummy dates so that the relative times between dates are retained.
Personal information (PI)	Subject-level data that can be linked to a data subject directly or indirectly, by reference to details such as name, identification number, location data or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of the subject.
Prosecutor risk	The risk of an adversary (individual or organisation) intentionally attempting to identify a data subject within a dataset. The adversary knows that a specific individual is in the dataset.
Pseudonymisation	A type of de-identification that both removes the association with a data subject and adds an association between a set of characteristics relating to the data subject and one or more pseudonyms. Typically, pseudonymisation is implemented by replacing direct identifiers (e.g. a name, a subject ID) with a randomly generated value.
Quasi (or indirect) identifier	Data which, in connection with other information, can be used to identify an individual with high probability, e.g. age at baseline, race, gender, medical information, events, specific findings, location.
Re-identification	Re-establishment of the association between a set of identifying data and the data subject found in data or documents.
Re-identification risk	The probability that re-identification could occur.
Reference population	The group of individuals who represent the basis for assessing the risk of re-identification. This group could be represented by the study population or by a larger group of individuals.
Risk threshold	The maximum amount of acceptable re-identification risk remaining in documents and data after an anonymisation process has been applied.
Secondary use	Uses and disclosures that are different from the purpose(s) for which the data was collected, as described in a clinical trial protocol and informed consent form.
Sensitive information	Any data which, in the event of re-identification, could be considered harmful for a data subject in terms of employability, reputation, insurability, self-esteem or stigma, or could result in loss of income. The perception of information as sensitive is subjective and examples include substance abuse, mental disorders and abortion.
Single out	To isolate some or all records that identify a data subject in the dataset by observing a set of characteristics known to uniquely describe that data subject.
Suppression	Removing certain data fields or entire records containing personal information.[50]

References

[1] Center for Global Clinical Research Data, Vivli. (Link)

[2] Clinical Study Data Request portal. (Link)

[3] Discussion Paper: Building a National Strategy for High-Cost Drugs for Rare Diseases: A Discussion Paper for Engaging Canadians. Available from (Link)

[4] Framework for Drugs for Rare Diseases. Available from (Link)

[5] Development of medicines for rare diseases. Available from (Link)

[6] Rare Diseases at FDA. Available from (Link)

[7] Nguengang Wakap S, Lambert DM, Olry A et al. (2020). Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet, 28(2):165–173. doi: 10.1038/s41431-019-0508-0

[8] CADTH Health Technology Review: Drugs for Rare Diseases: A Review of National and International Health Technology Assessment Agencies and Public Payers' Decision-Making Processes. Available from (Link)

[9] Smith CIE, Bergman P, Hagey DW. (2022). Estimating the number of diseases - the concept of rare, ultra-rare, and hyper rare. iScience, 25(8):104698. doi: 10.1016/j.isci.2022.104698.

[10] Echeverry-Quiceno LM, Candelo E, Gómez E et al. (2023). Population-specific facial traits and diagnosis accuracy of genetic and rare diseases in an admixed Colombian population. Sci Rep, 13(1):6869. doi: 10.1038/s41598-023-33374-x.

[11] Clinical Trial Data Sharing Ecosystem. Available from (Link)

[12] https://chdifoundation.org/

[13] External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use. Available from (Link)

[14] Guidance document on Public Release of Clinical Information. Available from (Link)

[15] Thorogood, A. (2020). International Data Sharing and Rare Disease: The Importance of Ethics and Patient Involvement. In: He Wu, Z. (Ed.), Rare Diseases. IntechOpen. Available from (Link)

[16] El Emam K. (2013). *Guide to the De-Identification of Personal Health Information.* Boca Raton, FL: CRC Press.

[17] Dunbar RIM (1992). Neocortex size as a constraint on group size in primates. Journal of Human Evolution, 22(6), 469–493. https://doi.org/10.1016/0047-2484(92)90081-J

[18] PHUSE De-Identification Working Group. Available from De-identification Standard for CDISC SDTM 3.2 Version 1.01.xls (live.com)

[19] Clinical data sharing: a proposed methodology to enable data privacy while improving secondary use. August 2023. Available from (Link)

[20] El Emam K, Dankar FK. (2008). Protecting privacy using k-anonymity. J Am Med Inform Assoc., 15(5):627–37. doi: 10.1197/jamia.M2716.

[21] Kniola L. Plausible Adversaries in Re-Identification Risk Assessment. Available from (Link)

[22] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. (2007). l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1):3-es.

[23] Bernier A. (2020). Rare disease data stewardship in Canada. FACETS, 5;5(1):836–63.

[24] Stanarević Katavić S. (2019). Health information behaviour of rare disease patients: seeking, finding and sharing health information. Health Information & Libraries Journal, 36(4), 341–356.

[25] Thorogood A. (2020). International Data Sharing and Rare Disease: The Importance of Ethics and Patient Involvement. Rare Diseases. Available from (Link)

[26] Titgemeyer SC, Schaaf CP. (2022). Facebook Support Groups for Pediatric Rare Diseases: Cross-Sectional Study to Investigate Opportunities, Limitations, and Privacy Concerns. JMIR Pediatr Parent, 5(1):e31411, doi: 10.2196/31411

[27] Rubinstein YR, Groft SC, Bartek R et al. (2010). Creating a global rare disease patient registry linked to a rare diseases biorepository database: Rare Disease-HUB. Contemp Clin Trials, 31(5):394–404. https://doi. org/10.1016/j.cct.2010.06.007.

[28] Iyer AA, Barzilay JR, Tabor HK. (2020). Patient and family social media use surrounding a novel treatment for a rare genetic disease: a qualitative interview study. Genetics in Medicine, 22(11):1830–7.

[29] Peer-to-peer Health Care. Available from (Link)

[30] Jacobs R, Boyd L, Brennan K et al. (2016). The importance of social media for patients and families affected by congenital anomalies: A Facebook cross-sectional analysis and user survey. J Pediatr Surg, 51(11):1766–1771. doi: 10.1016/j.jpedsurg.2016.07.008.

[31] Nabity-Grover T, Cheung CMK, Thatcher JB. (2020). Inside out and outside in: How the COVID-19 pandemic affects self-disclosure on social media. Int J Inf Manage, 55:102188. doi: 10.1016/j.ijinfomgt.2020.102188.

[32] Kordzadeh N, Warren J. (2017). Communicating personal health information in virtual health communities: An integration of privacy calculus model and affective commitment. Journal of the Association for Information Systems, 18(1):45–81.

[33] PHUSE Terminology Harmonisation in Data Sharing and Disclosure Deliverables. Available from (Link)

[34] TransCelerate. (2023). Clinical Data Sharing: A Proposed Methodology to Enable Data Privacy While Improving Secondary Use. Available from (Link)

[35] Disclosure of Adverse Events in Anonymized Documents. PHUSE presentation. Available from (Link)

[36] Parker R, Aggleton P. (2007). HIV and AIDS-related stigma and discrimination: a conceptual framework and implications for action. In: Culture, Society and Sexuality, pp. 459–474. Routledge.

[37] von der Lippe C, Diesen PS, Feragen KB. (2017). Living with a rare disorder: a systematic review of the qualitative literature. Molecular Genetics & Genomic Medicine, 5(6):758–73.

[38] Thomas M, Mackes N, Preuss-Dodhy A et al. (2024). Assessing Privacy Vulnerabilities in Genetic Data Sets: Scoping Review. JMIR Bioinformatics and Biotechnology, 5:e54332.

[39] Takashima K, Maru Y, Mori S et al. (2018). Ethical concerns on sharing genomic data including patients' family members. BMC Med Ethics 19, 61. https://doi.org/10.1186/s12910-018-0310-5

[40] Jiang A, Handley RR, Lehnert K, Snell RG. From pathogenesis to therapeutics: a review of 150 years of Huntington's disease research. International Journal of Molecular Sciences, 24(16):13021.

[41] Home - Clinical Data Publication - clinicaldata.ema.europa.eu

[42] El Emam, K. (2013). *Guide to the De-Identification of Personal Health Information.* CRC Press.

[43] Kussel T, Brenner T, Tremper G et al. (2022). Record linkage based patient intersection cardinality for rare disease studies using Mainzelliste and secure multi-party computation. Journal of Translational Medicine, 20(1):458.

[44] Bamford S, Lyons S, Arbuckle L, Chetelat P. (2022). Sharing Anonymized and Functionally Effective (SAFE) Data Standard for Safely Sharing Rich Clinical Trial Data.

[45] Raising the voice for rare diseases: under the spotlight for equity. (2023). eClinicalMedicine, 57, 101941.

[46] Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine. (2015). Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington (DC): National Academies Press (US). Appendix B: Concepts and Methods for De-identifying Clinical Trial Data. Available from https://www.ncbi.nlm.nih.gov/books/NBK285994/

[47] Dyke SO, Linden M, Lappalainen I et al. (2018). Registered access: authorizing data access. European Journal of Human Genetics, 26(12):1721–31.

[48] El Emam K, Dankar FK, Neisa A, Jonker E. (2013). Evaluating the risk of patient re-identification from adverse drug event reports. BMC Medical Informatics and Decision Making, 13:114.

[49] Article 29 Data Protection Working Party Opinion 05/2014 on Anonymisation Techniques. Available from Link

[50] Kohlmayer F, Prasser F, Kuhn KA. (2015). The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss. Journal of Biomedical Informatics, 58:37–48.