PP13 Rare disease data – overcoming barriers to controlled access sharing

PHUSE Transparency Working Group

Lead Authors: Helen Spotswood
Karolina Stepniak

Rare disease clinical trial data are not routinely shared, frequently falling outside of standard company transparency policy. PHUSE Rare Disease Clinical Data Sharing work stream has developed a whitepaper to review potential barriers to data sharing, e.g., risk of re-identification and invasion of privacy and to provide recommendations to encourage the sharing of rare disease data with the research community. This poster summarises some of the content with a focus on recommendations

The **definition of Rare Disease varies across jurisdictions**, and there is no single definition that is accepted globally

EMA
Health Canada
<5 per 10,000

FDA
***<8.6** per 10,000

*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 based on population at that time)
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 2 - 3 in 100, 000

Rare Diseases should be recognized as presenting a spectrum of risk, opening the door for flexible policy
-term 'Rare Disease' covers a **broad prevalence spectrum**
-considerable **variation in the sensitivity of the data** (e.g. Neurofibromatosis type 1 (NF1) with facial dysmorphology associated with stigma), many rare diseases are without physical manifestations (e.g. Acute Myeloid Leukemia)

High **RISK** Low

Factors that could increase participant re-identification risk considered in the context of a controlled access platform

Low prevalence?

Sensitive Information and Identifying disease traits?

Not a well studied disease area (low clinical trial population)?

Genomic data?

Single site studies?

Other information in public domain?

Social media activity?

If a data contributor uses **prosecutor risk model as standard** for controlled access data sharing

- Selection of the single study population as the reference is the most straightforward, least resource intensive and conservative option
- **May impact data utility especially for RD settings (small studies) to extent it will not support the research**

If a data contributor has an established **journalist risk model** for controlled access data sharing

- External reference population could be considered for RD but may require a manual review instead of semi-automated from clinical trial registries
- Consider pooling of studies in a development program for risk quantification even if only one study is to be shared

Handling outliers that drive information loss:
-Participant suppression: ensure that comparison to anonymized documents does not allow inferences to be drawn about the suppressed participants
-**Multimodal k-anonymity risk analysis** may be applied with greater data utility for a major subpopulation in the dataset at the same time as sacrificing data utility in minor subpopulation(s)

Research project's focus is important - maximizing information on outliers may be preferable; approach needs to be aligned with research proposal

Due diligence process carefully managed by company transparency teams
- embracing opportunity to connect with the data requestor while using platform such as Vivli
- before data preparation - a deep understanding of the disease area, the data and any indirect identifiers/sensitive data will be needed, and the requirements of the research proposal
- extra controls may be warranted e.g. contractual agreement by all researchers accessing the data and working on a publication

One transparency vison - Maintaining the highest possible alignment towards patient privacy protection across various clinical trial transparency initiatives
- o a review of previously released information should inform the anonymization approach

Conclusions

Model assumptions should not by default be made more restrictive

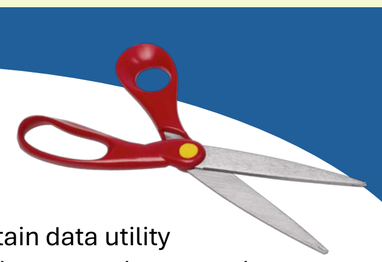
- RD at the upper end of the prevalence spectrum/ well-studied may be sharable with limited adaptation of company standard approach
- With small study size, small population and data sensitivity; company standard approaches
- may not result in a favorable privacy risk /data utility balance for RD

Rare disease data sharing will often be more resource intensive (tailored approach)

- A bespoke anonymization approach may be needed
  - Reduction of sensitivity
  - Prioritisation of data utility tailored to specific request
- providers may need to prioritize data requests; quality and value assessment - resources are not limitless.

Decisions on a case-by-case basis for rare diseases
- o companies may choose to not proactively list rare disease datasets but anonymize and share the data only with a specific request in mind

Systematic treatment of public disclosures of RD data with stricter risk thresholds, resulting in more stringent transformations, may create potential disclosure biases, whereby only releases of non-RD data may retain any meaningful utility



Data minimization is an important tool to maintain data utility
- Removal of indirect identifiers not required for the research proposal can allow retention of more granular information in other required variables.

- Complete or partial exclusion of some datasets should be considered, especially those with sensitive information.
  -A reduction in sensitivity could allow use of less conservative model options and reduce resource requirements

increasing sensitivity

larger acceptable risk threshold

EMA Policy 0070 & Health Canada PRCI have established a re-identification maximum risk threshold of 0.09 for public release, which is quoted as 'conservative'

Risk Threshold for maximum overall risk of re-identification should consider the sensitivity of the data

The acceptable maximum risk of re-identification in the data itself if an attack where to occur should be aligned with the context of where the data will be shared (e.g., public versus controlled access). This means that a smaller equivalence class to meet the same overall risk of re-identification may be acceptable where there are strong security and privacy practices and low motivation to attempt to re-identify

If **genomic data** is not required for a research purpose, it should be considered for exclusion
- privacy risk will vary e.g., somatic vs germline variation, frequency of SNPs and correlation in the population
- transparency experts may not be experts in genomic data types contribution to identifiability
- internal experts on these data types may be needed to be consulted e.g., on frequency of variants and their combinations to make considered decisions.

To access the draft Whitepaper, see here