PhUSE 2017

Paper PP10

ANONYMISATION OF CLINICAL STUDY REPORTS
A medical writer's perspective

Marie-Anne Thil, Keyrus Biopharma, Belgium
Julie De Wever, Keyrus Biopharma, Belgium

ABSTRACT

Complying with Policy 0070 entails multiple challenges. Clinical study reports must be anonymised for submission under this policy to prevent patients and professionals who participated in clinical trials from being identified. In its guidance, the EMA proposes three approaches for the anonymisation of CSRs: masking, randomization or generalisation. As a contract research organization, Keyrus Biopharma is developing a tool called KeyDAN to provide an adequate and efficient solution for CSR anonymisation. It is intended that this tool will provide the most appropriate way to increase transparency and to make available individual patient data comply with privacy and data protection laws and thus avoidthe risk of subject re-identification. In this paper, we present our perception of the different anonymisation approaches with strengths and weaknesses and also the challenges we faced in the development of this specific tool.

CONTEXT

For the past 25 years, legal requirements have promoted the pharmaceutical industry's disclosure of clinical trial results. In order to provide and extend an environment of transparency, one of the goals of the EMA is to see the publication of clinical data once a decision has been reached on the EU-wide marketing authorisation. The most recent step in this direction was the publication of Policy 0070 (EU region) which implements policy in two phases: Phase 1 came into force in January 2015 and concerns the public disclosure of clinical reports, and Phase 2, to be implemented at a future date, concerns the publication of individual patient data[1]. This paper discusses the first phase.

Public disclosure is not just a legal requirement; it is also an ethical responsibility and each company should commit to disclose its clinical trial data. This transparency is a must-have for improving public trust in the pharmaceutical industry. Public disclosure is also important for the scientific community by avoiding duplication of work, fostering innovation, allowing data re-assessment and data re-use and therefore might become a key asset for accelerating science.

Despite these positive aspects, there is a looming concern about patients' privacy and how to ensure both a sufficient protection and utility of the data. As a matter of fact, according to Policy 0070 and the associated external guidance[1,2], clinical data (clinical reports and individual patient data) related to trials should be submitted in a context of:

- A marketing authorisation application,
- A procedure under Article 58 of Regulation (EC) N<u>o</u> 726/2004
- or a new indication/line extension applications for authorised medicinal products need to be disclosed[2].

The term "clinical report" refers here to the clinical overviews and clinical summaries of the common technical document (CTD modules 2.5 and 2.7) and the clinical study reports ([CSR] included in CTD module 5) as well as some of its appendices such as the clinical study protocol and amendments, the case report form and the statistical analysis plan. These documents contain significant amounts of sensitive personal data, which cannot be disclosed as they are. Therefore, in addition to the primary clinical report developed for the competent authorities, a secondary document now needs to be issued for public disclosure – the secondary, or redacted, report. In this new document, all sensitive data (personal patient data [PPD] and company confidential information [CCI]) should be removed to avoid identification of individual patients. This process is called **anonymisation**. Each action taken in that direction inevitably impacts the usefulness of the data. The trick is then to keep useful data while ensuring the confidentiality of individual personal data (i.e. guaranteeing a low risk of re-identification either fortuitous or following a re-identification attack [see definition box]).

The EMA has published the external guidance[2] in which effective anonymisation has been defined by the three following criteria:

- No possibility to single out an individual,
- No possibility to link records relating to an individual,
- No information can be inferred concerning an individual.

When one or more criteria are not met, the risk of re-identification (see definition box) must also be assessed and shared along with the anonymisation methods in an "**anonymisation report**" (which will be part of the disclosure package submitted to the EMA.

In this paper, our focus is mainly on the CSR which consists of aggregated data and could therefore be considered as a pseudo-anonymised document (see definition box). But several sections of the CSR, such as safety narratives, individual profile in efficacy or pharmacokinetic results, or protocol deviations, describe unique, detailed events and are considered high-risk sections.

ANONYMISATION METHODS

In addition to the EMA guidance[2], organisations of biopharmaceutical companies (Pharmaceutical Researchers and Manufacturers of America [PhRMA], European Federation of Pharmaceutical Industries and Associations [EFPIA] and TransCelerate Biopharma Inc.) have already worked on this issue and have provided some valuable insights while keeping in mind the need to have a balance between clinical data utility and the risk of re-identification[1,3,4].

Anonymisation of the CSR requires the determination of:

1. The direct and quasi-identifiers (Safe-Habor method of de-identification) [3],
2. The required level of anonymization,
3. The methodologies to be used.

More specifically, the CSR should be reviewed to identify any information which may directly, or when linked to other information, identify an individual (staff or patient/subject). Whether this information should be anonymised or not might also depend on the risk level of the study. Indeed, the study characteristics such as its population (number of patients, specificities such as age, race or gender), the pathology studied (rare diseases or not) or the geographical location of the study (countries/regions, number of centres) permit the definition of a risk ranging from high (for instance, small specific population suffering from a rare disease and studied in one study centre in a very specific geographical area) to low (for instance, large population suffering from a very common disease in numerous study centres worldwide)[5]. These characteristics define not only the level of anonymisation but also the methods/combination of methods used to anonymise the report.

Methods described in the EMA external guidance include:

1. Removal/masking,
2. Generalization and,
3. Replacement/randomisation.

Removal/masking consists of masking the data and text to be removed with a black box (which should identify if the redacted text is CCI or PPD). It appears to be the easiest method compared to generalisation where data should be aggregated (using ranges and categories) or replacement/randomisation where data are replaced by new ones which can be randomly attributed. Therefore it is the best solution when reports have been written before the policy came into application. This reactive data anonymisation is progressively being replaced by a proactive approach using more replacement/randomisation techniques. These methods offer various advantages and disadvantages which should be considered on a case by case basis. Experience and hindsight will be of the utmost importance for the future choice of anonymisation method in each specific clinical report (Table 1).

Masking undoubtedly presents a low risk of re-identification but parts of data utility are lost in this process. Besides, as long as the process is performed manually, there is a non-negligible risk of error or to miss some data to be anonymised. In this case, potential risk that a person could be re-identified still exists. This method could, at least partially, be automated thus increasing its reliability. But a manual quality check would still be necessary. Generalisation or aggregating data preserves data utility to some extent but special attention must be paid to the ability to link the data. Replacement or recoding of the data with new random data (with irreversible destruction of the key code) allows creation of a document which seems analysable. However, the replaced content may lead to incorrect assumptions or lead to uncertainty[5]. Whatever the method chosen, thorough quality checks must be undertaken to keep re-identification risk and loss of data utility to a minimum.

Table 1: Advantages and disadvantages of anonymisation methods

Reactive (retrospective) data anonymisation
Proactive data anonymisation

	Removal / Masking	Generalisation	Randomisation/Replacement
Data utility*	Low	Better	Medium (?)
Risk of re-identification*	Low	Medium Ability to link records to be assessed	Medium (?)
Reliability (risk of errors or "missing")*	Medium to high	Medium to low	Low
IT tools/ automated methods	Could partially be done by IT tools	Could partially be done by IT tools	Yes

** Impact of anonymisation method is strongly dependent on the study specificities and study risk level*

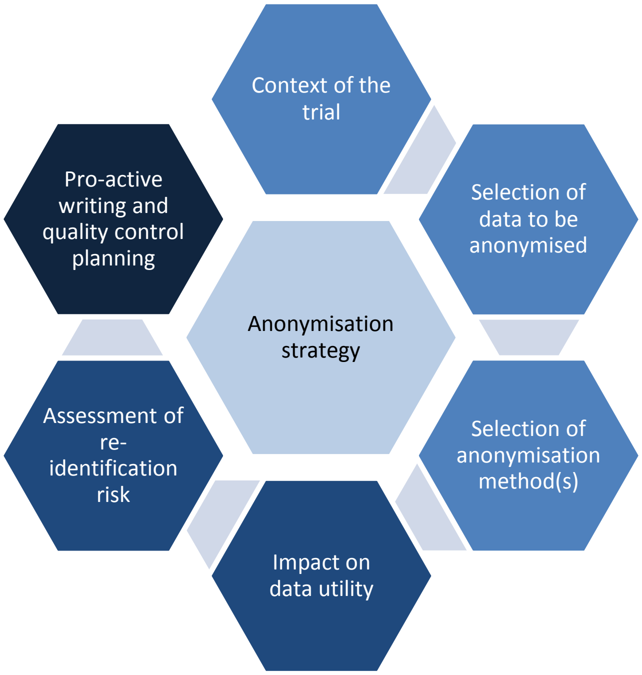
CHALLENGES

It is clear that a combination of all methods is probably the best approach bearing in mind the differences in risk of the different CSR sections and of each variable category (such as date, patient's identifiers, quantitative and qualitative variables). The **anonymisation strategy** or anonymisation plan must carefully take into account the specificities of the conducted studies (Figure 1). Indeed, each study is characterised by its context i.e. the size of its population, the country and number of centres, the studied pathologies (particular attention should be paid to rare diseases). As mentioned earlier, these characteristics define the risk level of the study. Taking this into account, the data to be anonymised should be carefully defined as well as the anonymisation method for each. Then, the impact on data utility should be assessed and the risk of re-identification should be evaluated. In particular, the threats posed by data mining or data linkage should be seriously considered as a risk evolving with technology (see definition box).

Anonymisation should therefore be anticipated and planned. Pro-active medical writing can be part of the effort. Indeed, while writing the primary clinical report for the competent authorities, it could be foreseen to avoid unnecessary information and already use generalisation (date, means, ranges) or aggregation methods (for instance to report outliers). Likewise, a first identification of data which may need anonymisation in the secondary report can be performed to ease automated masking. This will save time and effort in the forthcoming redaction (by decreasing the amount of data to be anonymised or flagging them).

Last but not least, a solid quality check needs to be put in place. Several questions may therefore arise: should a 100% quality check be performed or could a partial check be considered? Should legal representatives and/or patient committees be involved?

Figure 1: Anonymisation strategy


Context of the trial
Selection of data to be anonymised
Selection of anonymisation method(s)
Impact on data utility
Assessment of re-identification risk
Pro-active writing and quality control planning
Anonymisation strategy

CONCLUSION

The usefulness of the data and risk of re-identification are key parameters in the anonymisation of clinical reports - usefulness should be maximised and risk minimised. It has been the case that reactive anonymisation mainly used removal or masking methods. However, as we go forward, considering the issue proactively, anonymisation methods could include a combination of methods. IT tools for automated masking or for replacement/randomisation of data are becoming essential and we have seen in the development of Keyrus Biopharma's tool, KeyDAN, that the input of a medical writer's perspective has proven to be both necessary and effective. However, this challenging topic will need constant fine-tuning by a multi-disciplinary team and anticipation and pro-active writing may be key assets for that process.

Definitions

Aggregated data are statistical data about several individuals that have been combined to show general trends or values without identifying individuals within the data.

Anonymisation is the process of turning data into a form that does not identify individuals and where identification is unlikely to take place[2]. It allows for a much wider use of information.

Company Confidential Information CCI is any information contained in the clinical reports submitted to EMA by the applicant (or marketing authorization holder) which is not in the public domain or publicly available and where disclosure may undermine the legitimate economic interest of the applicant (or marketing authorization holder)[2].

Data linkage is a technique that involves bringing together and analysing data from a variety of sources, typically data that relates to the same individual[2].

Data mining is the activity of going through big data sets to look for relevant or pertinent information[2].

Direct identifiers are elements that permit direct recognition or communication with the corresponding individuals (e.g. name, address, phone numbers)[2]

Quasi-identifiers are variables representing an individual's background information that can indirectly identify individuals (e.g. initials, date of birth, gender)[2]

Pseudo-anonymisation consists of replacing one attribute (typically a unique attribute) in a record by another. When pseudo-anonymisation is used alone, the natural person is still likely to be identified indirectly[2]

Personal patient data (PPD) are any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to their physical, physiological, mental, economic, cultural or social identity[2].

Re-identification attack is a deliberate attempt to determine the identity of one or more individuals in a dataset or data base that is publicly share[2,6]

REFERENCES

1. Policy 0070 European Medicines Agency policy on publication of clinical data for medicinal products for human use. Published October 2014. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf. Accessed 21 September 2017.

2. External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal product for human use (Policy 070). Published April 2017. Accessed 21 September 2017.

3. Transcelerate – Data De-identification and Anonymisation of Individual Patient Data in Clinical Studies – A Model Approach. Published in 2013 http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/CDT-Data-Anonymization-Paper-FINAL.pdf Accessed 21 September 2017.

4. TransCelerate - Clinical Study Reports Approach to Protection of Personal Data. Published 2014 http://www.transceleratebiopharmainc.com/wp-content/uploads/2014/08/TransCelerate-CSR-Redaction-Approach.pdf. Accessed 21 September 2017

5. Transcelerate 2016 Protection of personal data in clinical documents – A model approach http://www.transceleratebiopharmainc.com/wp-content/uploads/2017/02/Protection-of-Personal-Data-in-Clinical-Documents.pdf Accessed 21 September 2017

6. El Emam K, Rodgers S, Malin B. 2015 BMJ. Anonymising and sharing individual patient data.2015 Mar 20;350:h1139. doi: 10.1136/bmj.h1139.

ACKNOWLEDGMENTS

The authors warmly thank Kirsten Dumaz and Neil Goodman for proofreading the poster and article, respectively, and Marie Guglielminetti for her support in preparing this paper.

Your comments and questions are valued and encouraged. Please contact the authors at:

Julie De Wever or Marie-Anne Thil
Keyrus Biopharma
Chaussée de Louvain 88
Lasne / 1380
Email: Julie.dewever@keyrus.be or marie-anne.thil@keyrus.be
Web: http://www.keyrusbiopharma.com