Preparing anonymisation reports in general and for an orphan drug in particular

Louise Martinsson
Swedish Orphan Biovitrum, Solna, Sweden

Correspondence to:
Louise Martinsson
Swedish Orphan Biovitrum AB
Tomtebodavägen 19 A, Solna
112 76 Stockholm, Sweden
+46 76 001 15 06
louise.martinsson@sobi.com

Abstract
In 2015, the EMA Policy 0070 came into effect as part of EMA's commitment to increased data transparency. In short, clinical reports included in regulatory applications for example, marketing authorisations are published on the EMA web page and thereby made publicly available. Before the clinical reports can be published, the applicant is required by legislation to protect personal data to ensure individual clinical study participants and other individuals involved in the study are not identified. The applicant has to describe how data protection of personal data has been ensured in an anonymisation report (AnR). This article describes the different steps necessary to prepare an AnR in general, a company's first experience of preparing an AnR for an orphan drug, and the key points learned from this experience.

EMA Policy 0070 (the Policy) came into effect in 2016.[1] The Policy is part of an EMA initiative to increase transparency of clinical data and applies to three regulatory procedures in the framework of the centralised procedure. Since many of the terms in the Policy are unfamiliar to many medical writers, a list of terms and definitions used in this article is included in Table 1.

The clinical reports included in, for example, marketing authorisation applications are made publicly available under the Policy on the EMA webpage.[2] As part of any application under the Policy, two new documents are required; a table on justifications of commercially confidential information (not covered by this article), and an anonymisation report (AnR). A preliminary analysis on practices of AnRs published up to December 31, 2017, is provided in an article by Billiones on page 22 in this issue of *Medical Writing*.[3]

Table 1. Terms and definitions

Anonymisation	The process of rendering data into a form that does not identify individuals and where identification is not likely to take place
Anonymised/de-identified data	Data in a form that does not identify individuals and where identification through its combination with other data is not likely to take place
Data	Data in the context of the Policy means characteristics or information, usually numerical, that are collected through observation. The word can also be used to describe statistics (i.e., aggregations or transformations of raw data).
De-identification	See anonymisation.
Direct identifiers	E.g., patient ID, patient name, patient address
Clinical reports	Clinical reports in the context of the Policy means the clinical overviews (submitted in module 2.5), clinical summaries (submitted in module 2.7), and the clinical study reports (submitted in module 5, "CSR") together with the following appendices to the CSRs: 16.1.1 (protocol and protocol amendments), 16.1.2 (sample case report form), and 16.1.9 (documentation of statistical methods)
Masking	An anonymisation technique in which data-identification data are irreversibly blocked
Personal data	"Personal data" shall mean 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 his physical, physiological, mental, economic, cultural or social identity.
Redaction package	The package contains the anonymised clinical reports included in the regulatory procedure under the Policy as well as some other documents defined in the Policy. A proposed redaction package is submitted first while the final redaction package is submitted after EMA's review.
Publishing	The act of making data publicly available
Re-identification	The process of analysing data or combining it with other data with the result that individuals become identifiable, sometimes also referred to as "de-anonymisation"
Re-identification attack	An attack to identify an individual participating in a clinical trial. The reasons for attempting an attack could be, for example, to identify a trial participant of special interest such as a famous actor or a politician or to embarrass the data controller or to undermine the public support for release of data (demonstration attack).
Risk	The probability of re-identifying a trial participant.
Quasi identifiers	E.g., age, geographical location, sex, age, race, ethnicity

Since the applicant is responsible by legislation to protect personal data that can lead to identification of an individual, the applicant has to ensure these data are anonymised. The purpose with the AnR is to describe:

- The methodology of the anonymisation technique applied by the applicant.
- The rationale for the methodology used.
- How the risks of re-identification of the personal data have been measured and managed.

Two different anonymisation methodologies can be used, a quantitative or a qualitative one.

This article describes the procedures needed to prepare an AnR in general and is based on the EMA template for AnRs (Annex 1.2 in the Policy guidance).[1] Figure 1 presents a flow chart of activities included in the AnR preparation. In addition, the steps taken during the authoring of the first AnR based on a quantitative methodology published on the EMA webpage[3] are described and the key points learned from this experience are shared.

The AnR presented in this article was prepared by Biogen, which was the marketing authorisation holder (MAH) for the medicinal product Alprolix® (indicated for the rare disease haemophilia B) in the United States, while Swedish Orphan Biovitrum AB (publ) (Sobi) as the MAH for Alprolix® in Europe, was responsible for submitting and revising the AnR after interactions with the EMA.

Preparing the anonymisation report

The headings in this section of the article are derived from the Policy AnR template (Annex 1.2 Section 1.2.2.1.2 in the Policy guidance)[1] and are also used in Figure 1.

Anonymisation methodology

As a first step, the applicant should choose if a quantitative or a qualitative methodology should be used to anonymise personal data. The EMA encourages using a quantitative approach although they accept a qualitative approach during the pilot phase of the Policy implementation (Chapter 3, Section 5.4.4 in the Policy guidance[1]). For the Alprolix® AnR, a quantitative methodology was chosen and the anonymisation technique masking was applied (Figure 2).

Recognising direct identifiers and quasi identifiers

As a second step, direct identifiers and quasi identifiers should to be identified. This has to be done independently whether a qualitative or a quantitative methodology is used. The Policy guidance[1] provides examples of direct identifiers and quasi identifiers. If there are no direct identifiers and no quasi identifiers, a different EMA AnR template should be used (Annex 1.13, in the Policy guidance).[1]

The direct identifiers and quasi identifiers used in the risk assessment for the Alprolix® AnR are presented in Table 2. Since all patients were male and no deaths were reported during the trial, sex and date of death were not considered as quasi identifiers. In addition to the direct identifiers and quasi identifiers, some data were considered to be extra sensitive, i.e., HIV, hepatic C status, and genotype. These sensitive identifiers

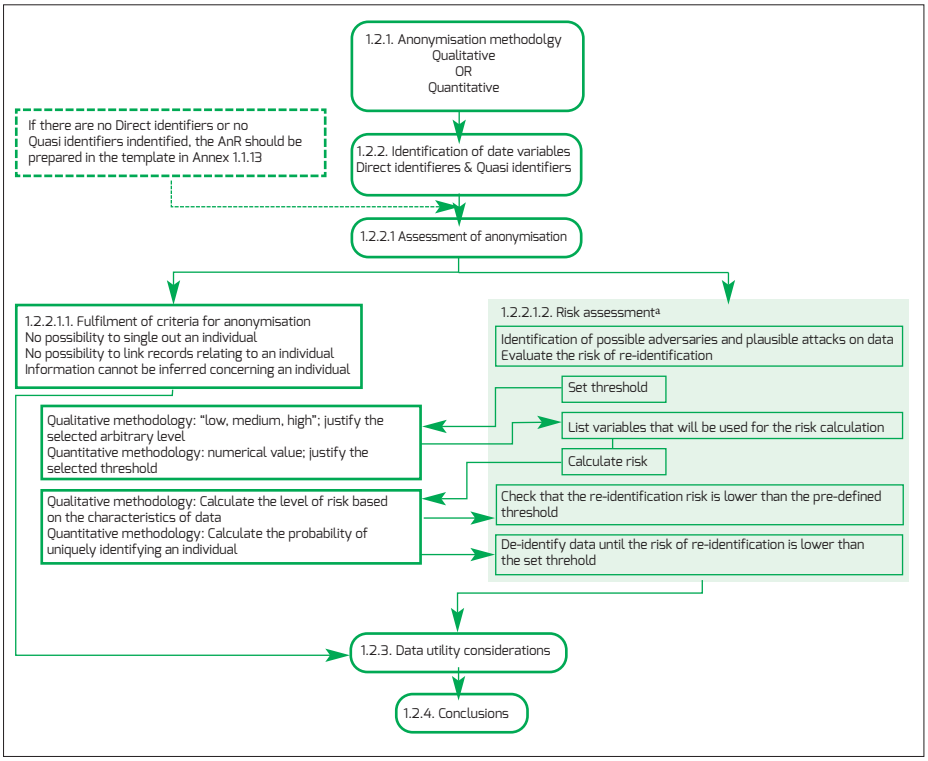

Figure 1. Flow chart of activities included in the preparation of anonymisation reports

[a] Step 1.2.2.1.1 is not needed if a risk assessment (see step 1.2.2.1.2) is performed.

The numbers indicate the section numbers in the EMA anonymisation report template (Annex 1.2, in the EMA Policy 0070 guidance).

Table 2. Direct identifiers and quasi identifiers in the Alprolix anonymisation report

Direct identifiers
Subject identifiers
Study site identifiers
Quasi identifiers
Age, birthdate
Race
Ethnicity
Country
Height, weight, BMI
Serious adverse events;
Adverse events of Interest relevant to haemophilia B and/or treatment[a,b]
Medical history[b]
Surgery details
Bleeding episodes
Calendar dates

Abbreviations: BMI, Body mass index.

[a] Development of inhibitors, incidence of allergic reactions, incidence of thrombotic events, incidence of suspected transmission of an infectious agent, infection events, adverse bleeding events.

[b] Verbatim text and preferred term.

were masked in the clinical reports but were not part of the risk assessment of re-identification.

Assessment of anonymisation

As a third step, an assessment should be performed to assess the extent of anonymisation needed to reduce the risk of re-identification. If the applicant can confirm or demonstrate that the following three criteria are fulfilled, the EMA AnR template (Annex 1.2, Section 1.2.2.1.2 in the Policy guidance)[1] should not be completed:

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

If one or more of the criteria are not met, the applicant should continue the assessment in the EMA AnR template (Annex 1.2 Section 1.2.2.1.2 in the Policy guidance).[1] For a qualitative methodology, there is no numerical threshold of risk of re-identification to be decided. Instead the applicant should use the arbitrary levels "low", "medium", and "high". The following definitions are examples of levels to be used:

- High risk: < 100 trial participants, rare disease.
- Medium risk: 100 to 1,000 trial participants.
- Low risk: > 1,000 trial participants.

If a *quantitative* methodology is used, the risk of re-identification and a numerical threshold of risk of re-identification should be decided. The EMA recommends to set the risk of re-identification to a maximum, i.e., 1, and the numerical threshold of risk of re-identification to 0.09. However, the EMA leaves it open to the applicant "to decide on the most appropriate threshold for public disclosure of clinical reports" as long as a justification of the selected threshold is provided.[1]

The risk of re-identifying personal data in the Alprolix® AnR was based on the combined trial population in all clinical reports included in the marketing authorisation application. A number of scenarios and iterations combining different quasi identifiers were performed as presented in Table 3. In the scenario presented in the last row of the Table, there were no trial participants with a unique value for any of the selected quasi identifiers. The risk with this scenario was 0.006 (1/67).

Data utility considerations

As a fourth step, the applicant should consider the data utility versus the re-identification risk.

Since haemophilia B is a rare disease it was considered necessary to mask all quasi identifiers and the sensitive data on an individual participant level, including full narratives, to protect the confidentiality of the trial participants even though this reduced the data utility. However, since aggregate summaries and analyses have the most scientific value and remained largely unmodified, Sobi still considered the remaining data as informative. This was accepted by the EMA even though they do

Note: 1) Change in ABR, consumption and number of injections is calculated as onstudy value - prestudy value.
2) Subjects [redacted] were excluded from the analysis because their pre-study regimen was sports prophylaxis. Subject [redacted] had a pre- and on-study ABR but not pre- and on-study consumption and number of injections. Subjects [redacted] had pre- and on-study consumption and number of injections but not a pre- and on-study ABR.

Figure 2. Example of anonymisation by masking

The blue box covers personal data that needs to be anonymised.

Table 3. Calculated risk for re-identification using different combinations of quasi identifiers in the Alprolix anonymisation report

Redacted quasi identifiers	Unredacted quasi identifiers	Subjects	Number of unique subjects[a]	Proportion of unique subjects (%)
Height, weight, BMI	Age, race, country, SAEs, surgeries, bleeding episodes	167	149	89.2
SAEs, surgeries, bleeding episodes, height, weight, BMI	Age, race, country	167	124	74.3
Age, race, country, height, weight, BMI	SAEs, surgeries, bleeding episodes	167	53	31.7
Race, country, SAEs, surgeries, bleeding episodes, height, weight, BMI	Age	167	20	12.0
Age, race, country, SAEs, Surgeries, bleeding episodes, height, weight, BMI	–	167	0	0.0

Abbreviations: BMI, body mass index; SAE, serious adverse event.

[a]Number of subjects for whom the combination of values in the un-redacted identifiers is unique.

not accept this approach by default (Chapter 2, Section 2.2 in the Policy guidance).[1]

Conclusion

As the fifth and final step, the applicant should declare that "the anonymisation report has been prepared following the guidance made available by EMA, and the anonymisation techniques have been applied consistently in the preparation of the documents comprising the Final Redacted Document package".

Key points learned from preparing an AnR

- Legal advice is important before choosing anonymisation methodology to ensure no data privacy laws are breached.
- If considering publishing personal data, have in mind that although a trial participant has consented to their data being published they have the right to withdraw their consent at any time.
- Statistical advice is crucial if a quantitative anonymisation methodology is chosen.
- Data anonymisation is a moving target as research, tools, and computational power evolve. Re-identification attacks (see Table 2) of anonymised data do occur and are becoming more common (Henriksen-Bulmer et al.,).[5]
- Note that anonymisation of personal data in relation to trial participants (Chapter 3, Section 5.3)[1] differ from personal data in relation to investigators, sponsors, and applicants (Chapter 3, Section 6).[1]
- During the "implementation phase" of the Policy, the EMA offers advice (telephone conferences, face to face meetings, written conversation). This service is provided to all applicants when submitting their first AnR under the Policy. The author´s experience was that EMA was interested in discussing the problems encountered, as well as the applicant's opinion of the Policy. The EMA also provided valuable feedback on the AnR and assisted in improving the quality of the AnR.
- The timelines in the Policy guidance did not apply when this article was authored. Check with the EMA when to submit your proposal package.
- Be sure to use the most recent version of the Policy guidance as it is being updated frequently. EMA Questions & Answers[6] provide useful tips on how to interpret the Policy guidance.[1]
- EMA offers small and medium sized companies a redaction tool licence for 12 months. An application for the tool should be done five months prior the expected CHMP opinion.

Concluding remarks

The preparation of the AnR is in its early stages and both the EMA and, in particular, applicants/MAHs have a steep learning curve ahead until the AnR can be considered a mainstream regulatory document. As the Policy, including the AnR, is still in the implementation phase and companies as well the EMA are still learning, the preparation of the Alprolix® AnR should only be consider as an example of how to prepare a quantitative AnR.

Acknowledgements

The author would like to thank Raquel Billiones and Maria Wikén Wintergren for their review of the manuscript.

Disclaimers

The opinions expressed in this article are the author's own and not necessarily shared by her employer or EMWA.

Conflicts of interest

The author Louise Martinsson is employed by Sobi, a biopharmaceutical company with a focus on rare diseases.

References

1. 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: http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_001799.jsp&mid=WC0b01ac0580b2f6ba.
2. European Medicines Agency website on Clinical data. Available from: https://clinicaldata.ema.europa.eu/web/cdp/home.
3. Billiones R. Anonymisation reports from 2016 to 2017: A preliminary analysis. Med Writ. 2018;27(4):22–6.
4. Henriksen-Bulmer J, Sheridan J. Re-identification attacks – A systematic literature review. IJIM. 2016;36:1184–92.
5. EMA Questions & Answers on the External Guidance of Policy 0070 on Clinical Data Publication (CDP). Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2017/04/WC500225881.pdf.

Author information

Louise Martinsson, PhD, has been a medical writer since 2007. She has been working at AstraZeneca and Linde Healthcare and currently holds a positions as senior medical writer at Sobi since 2016 where she is responsible for implementing EMA and FDA clinical data transparency regulations.