# Anonymisation reports from 2016 to 2017: A preliminary analysis

**Raquel Billiones**
Clinipace Worldwide, Zurich, Switzerland

## Correspondence to:
Raquel Billiones
Clinipace AG
Chriesbaumstrasse 2
CH-8604 Volketswil, Zurich, Switzerland
rbilliones@clinipace.com

## Abstract
The anonymisation report (AR) is a new and relatively unknown regulatory document, submitted as part of the redacted package of a marketing authorisation application under the EMA Policy 0070. The report documents the methodology of anonymisation in each package and the rationale for these methods. As of December 31, 2017, 64 ARs have been published on the clinical data website of the European Medicines Agency. A preliminary high-level analysis of these reports was performed, with the aim of gaining information on the current industry practices in anonymisation and AR preparation. After excluding 12 ARs from packages that did not contain protected personal data, 52 ARs were analysed. Information on anonymisation methodology, re-identification risk assessment, data utility assessment, and use of software is presented.

## Background
The EMA Policy 0070 (referred to henceforth as "the policy") version 1.0 was finalised on March 2, 2016. At the time of writing, the policy has been revised three times, most recently (version 1.3) on September 20, 2017. As a requirement of the policy, certain documents ("clinical reports") in the marketing authorisation applications (MAA) are to be published on the clinical data website of the European Medicines Agency (EMA) (https://clinicaldata.ema.europa.eu). However, before these reports are published on a portal that is available to the general public, anonymisation of protected personal data and commercially confidential information, if applicable, is necessary. The resulting anonymised dossier that will be published is called the "redacted document package"; the first redacted packages were published on October 20, 2016. The policy provides guidance on the anonymisation process but "is not intended to mandate any specific methodology but to highlight to applicants/marketing authorisation holders (MAHs) the available techniques and those that EMA considers most relevant in the context of the anonymisation, to ensure that clinical reports submitted to EMA for publication are rendered anonymous prior to publication."[1]

The "clinical reports" to be disclosed are documents (e.g., study reports, clinical summaries and overviews) that regulatory medical writers are very familiar with, except one. The anonymisation report (AR) is a new requirement under the policy and documents the methodology of anonymisation and assessment of re-identification risks in each package. Many companies are preparing the redacted packages and the AR for the first time and the industry is still gathering experience and know-how. The policy provides guidance on the structure and content of AR in Annex 1.2 Anonymisation Report – Template.

Approximately 14 months after the launch of the EMA clinical data website (October 2016 to December 2017), 64 redacted packages of marketing application procedures had been published, and each package contained an AR. This paper describes a high-level content analysis of these published ARs, with the purpose of gaining information on the current status and practices in anonymisation and the preparation of redacted packages as documented in the AR.

## Methods
The EMA clinical data website (https://clinicaldata.ema.europa.eu) was accessed under the academic and other non-commercial research purposes terms of use. Using the advance search option, all procedures of all types from September 2016 to December 2017 were retrieved (Figure 1) without the use of filters.

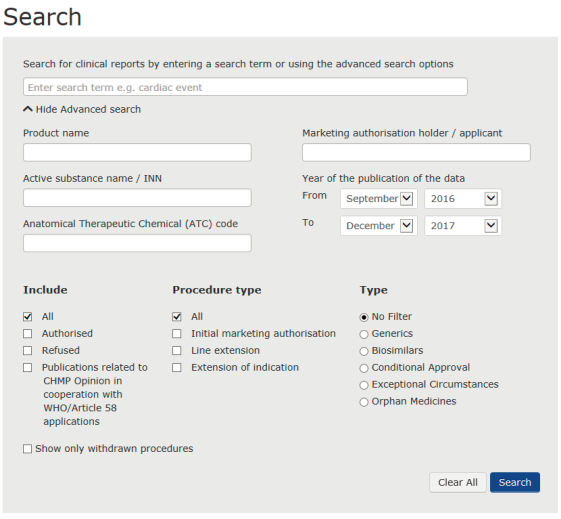


***Figure 1. Search method used to retrieve anonymisation reports from the EMA clinical website***
https://clinicaldata.ema.europa.eu
Screenshot is used with permission from the EMA.

The search results were exported into an Excel file. Each package was accessed, with particular focus on the AR. Each AR was downloaded for further scrutiny. In addition to the package results obtained in the Excel export file, information on the content of the ARs was extracted, focusing on the following:

- Option used to establish effective anonymisation.
- Method of risk assessment of patient re-identification.
- Anonymisation approach.
- Data utility assessment.
- Use of software.

Additional analysis on orphan drug applications (ODA) was also conducted, as the risk for re-identification of subjects would appear to be higher in rare disease research and small populations.

This paper focuses on ARs only; the full redacted packages were not analysed. The methodology of this analysis was not validated but deemed sufficient to provide descriptive information about the ARs analysed. The analysis did not take into account potential overlaps among the ARs due to indication and line extensions of the same product.

All screenshots and publicly available information shown in this article are used with the permission of the EMA.

## All anonymisation reports from 2016 to 2017

A total of 64 redacted packages submitted by 29 MAHs were published from October 20, 2016, to December 31, 2017; 64 ARs in these packages were examined. Twelve packages did not contain any protected personal data; their ARs were excluded, leaving a total of 52 ARs for further analysis. The number of pages of these 52 ARs ranged from 4 to 53 pages. Figure 2 shows the individual AR, the month of publication, and other information.

A summary of information of anonymisation methodology in the 52 reports is provided in Table 1.

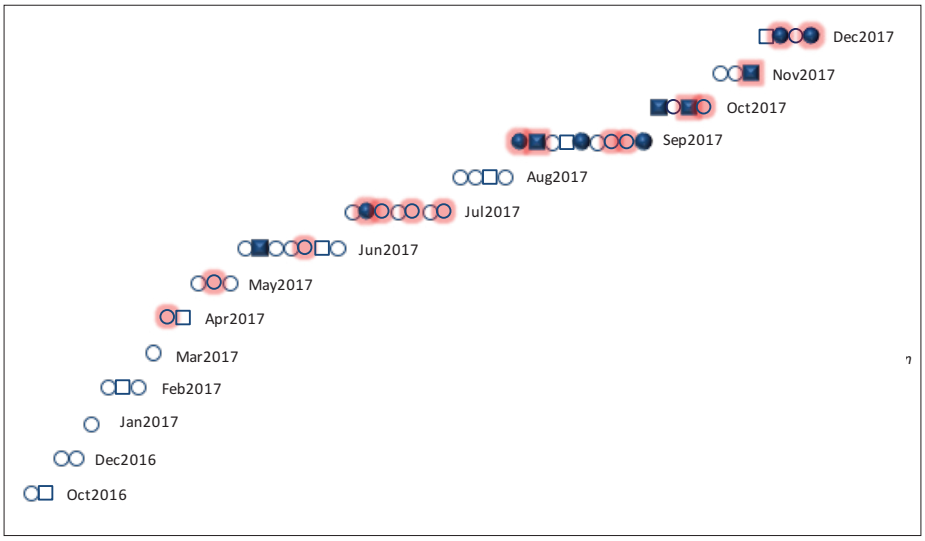


***Figure 2. Anonymisation reports published on the EMA clinical data website from October 2016 to December 2017 (n=52)***

Each symbol represents 1 AR.

Circles: ARs for non-orphan MAA (*n*=40)

Squares: ARs for MAA with orphan drug designation (*n*=12)

Filled symbols: ARs for MAA with quantitative risk assessment (*n*=11)

Symbols with red shadow: ARs for MAA that used automated redaction tools or artificial intelligence systems (*n*=16)

***Table 1. Summary of methods described in anonymisation reports, October 2016 to December 2017***

| Information | All Reports *N*=52 | | Orphan Drug Application Reports *N*=12 | |
|---|---|---|---|---|
| | *n* | % | *n* | % |
| **Demonstration of effective anonymisation** | **52** | **100%** | **12** | **100%** |
| Option 1: prevent singling out, linkage, inference | 5 | 9.62% | 2 | 16.67% |
| Option 2: re-identification risk assessment | 45 | 86.54% | 10 | 83.33% |
| Options 1 and 2 combined in 1 AR | 2 | 3.85% | – | – |
| **Risk assessment of re-identification method** | **52** | **100%** | **12** | **100%** |
| Qualitative risk assessment (low, medium, high) | 39 | 75.00% | 6 | 50.00% |
| Quantitative risk assessment (numerical threshold) | 8 | 15.38% | 4 | 33.33% |
| Not applicable[a] | 5 | 9.62% | 2 | 16.67% |
| **Anonymisation method** | **52** | **100%** | **12** | **100%** |
| Non-analytical | 32 | 61.54% | 4 | 33.33% |
| Analytical | 11 | 21.15% | 5 | 41.67% |
| Not clearly specified | 9 | 17.31% | 3 | 25.00% |

[a] All 5 ARs that used option 1

**Effective anonymisation**

The policy provides two options to establish effective anonymisation as described in Section 3.2.1 Anonymisation Criteria. The first option is to demonstrate that the anonymisation method used removes the possibility of singling out, linkage to, and inference to an individual patient. The second option is the evaluation of re-identification risk and demonstrating that the anonymisation method used mitigated this risk to the lowest level.

Only five ARs used the first option as shown in Table 1. An overwhelming majority (*n*=45; 86.54%) of ARs used the second option and assessed the risk of re-identification. For these ARs, the section on Fulfilment of the Criteria for Anonymisation was marked as "not applicable". Two ARs by the same MAH used both options.

**Qualitative vs quantitative assessment method of re-identification risk**

The methods of assessing the risk of patient re-identification are presented in Table 1.

The five MAHs that used only effective anonymisation criteria option 1 did not perform re-identification risk assessment.

Of the 45 ARs that used effective anonymisation option 2, 39 (75%) assessed the risk of re-identification in a qualitative manner, using the scale of low, medium, or high risk. The scale is arbitrary and not detailed in the policy but most MAHs attempted to define this in their ARs.

Eleven ARs assessed the risk quantitatively, i.e., by calculating the probability of re-identification and measuring the risk numerically. The policy recommends a conservative threshold of 0.09 but allows for another threshold to be used as long as this is appropriately justified. There is a trend towards more frequent use of quantitative risk assessment towards the end of 2017 as shown in Figure 2.

**Analytical vs non-analytical anonymisation approach**

Most ARs (32 [61.54%]) described their anonymisation method as non-analytical, whereas 11 (21.15%) ARs claimed using the analytical method of anonymisation (Table 1).

> There is a trend towards the use of quantitative re-identification risk assessment and automated or artificial intelligence systems as the industry develops new tools and gains experience.

The terms "analytical" and "non-analytical" are not clearly defined in the policy and were used in the AR rather ambiguously, possibly coming from the policy's Section 5.4 Anonymisation Process: "Applicants/MAH may *not* follow, in an initial phase, an *analytical approach*, and therefore it will not be necessary to calculate the risk of re-identification. In such cases *step 4* of the anonymisation process could be omitted."[1]

Step 4 refers to the determination of quantitative risk of re-identification threshold. Hence, it is justifiable that many MAHs used the terms "non-analytical" and "qualitative" synonymously to refer to their anonymisation methodology. However, there are a few ARs that did not equate qualitative risk assessment with non-analytical approach, as demonstrated in this excerpt from one report: "Assessment of anonymisation has been performed using an *analytical* approach that evaluates criteria for anonymisation and expected risk factors on a *qualitative basis.*"[2]

Several MAHs described their "non-analytical" anonymisation approach as reviewing the documents manually and deciding on the text and numbers to be redacted based on predefined criteria. However, there were also ARs that used a "non-analytical" anonymisation approach using automated redaction tools.

All 11 ARs that used the quantitative risk assessment method also described performing anonymisation in an analytical manner. After anonymisation, the risks of re-identification were <0.09, the threshold suggested by the policy.

**Use of software**

Table 2 summarises the information on software mentioned in the ARs.

The use of some form of software was specified in 32 ARs, 16 of which referred to automated redaction tools or artificial intelligence systems (see Figure 2). Six ARs by two MAHs used the Lexicon Tool Suite by Privacy Analytics; 10 did not specify the proprietary name of the software used.

Sixteen ARs mentioned using manual redaction tools but only six specifically mentioned the proprietary name (Adobe Acrobat/Nuance). Twenty ARs did not mention the use of any kind of software.

The use of Lexicon Tools Suite allowed the

***Table 2. Summary of software use described in anonymisation reports, October 2016 to December 2017***

| Information | All Reports *N*=52 | | Orphan Drug Application Reports *N*=12 | |
|---|---|---|---|---|
| | *n* | % | *n* | % |
| ARs that do not mention the use of software | 20 | 38.46% | 4 | 33.33% |
| ARs that mention the use of software | 32 | 61.54% | 8 | 66.67% |
| ARs that used automated redaction tool/ artificial intelligence (AI) system | 16 | 30.77% | 3 | 25.00% |
| Lexicon Tool Suite | 6 | 11.54% | 3 | 25.00% |
| unspecified automated redaction tool/ AI system | 10 | 19.23% | – | – |
| ARs that used manual redaction tool | 16 | 30.77% | 5 | 41.67% |
| Adobe Acrobat/Nuance | 6 | 11.54% | 1 | 8.33% |
| unspecified manual redaction tool | 10 | 19.23% | 4 | 33.33% |

MAH to pseudonymise personal data by the use of transformation or recoding algorithms; redaction was only done where transformation was not possible.

Redaction of published patient data listings was performed in three procedures by one MAH that used Lexicon Tool Suite. It is to be noted that disclosure of patient listings is currently not obligatory, being out of scope in the phase 1 implementation of the policy. However, publication of listings is planned for phase 2 of the policy implementation.

**Data utility considerations**

The policy considers the impact of data transformations and redactions on the scientific utility of the report. In principle, a balance between an acceptably low risk of re-identification and maintainance of data utility should be achieved. The majority of ARs discussed data utility considerations descriptively. Only those ARs that used Lexicon Tool Suites ($n$=6) described using metrics to assess data utility post-anonymisation. In these ARs, data utility was assessed by a) precision metric that measures data distortion following anonymisation b) subjective assessment criteria pertaining to the accuracy of analysis results based on the assumption that a secondary data user is planning to replicate the original results of the trial. Based on these metrics, data transformation combined with redaction resulted in higher data utility compared to a redaction-only approach.

**Anonymisation reports for orphan drug applications**

Of the 52 ARs analysed, 12 were from orphan drug applications (ODA) as summarised in Tables 1 and 2 and shown in Figure 2. These applications are of special interest as they deal with rare diseases and studies with small sample sizes. Six of the ODA ARs used the qualitative risk assessment method, four used the quantitative method, whereas two used the first option to establish effective anonymisation and hence did not perform any risk assessment. Five ODA ARs used the analytical approach of anonymisation but only four performed quantitative risk assessment. Three ARs used the Lexicon Tool Suite for data transformation, automated redaction and data utility metrics.

The first AR that documented a quantitative measure of re-identification risk was that of an orphan drug, published in June 2017 (see Figure 2). The lessons learned from this redacted package are described in the article by Martinsson on page 27. There is no indication that ODAs are more likely to use the qualitative approach of risk assessment due to small study populations.

## General observations and recommendations

**Analytical vs non-analytical approach to anonymisation**

The terms "analytical" and "non-analytical" were frequently used in the ARs but rather ambiguously as described above. It is suggested that MAHs should provide clear definitions when using these terms in the ARs.

**One-size-fits-all anonymisation methodology**

Several ARs described one anonymisation methodology that appeared to apply to all studies in the package. While this may be true in some cases, this should not be the general practice. MAAs usually consist of trials of different phases and sample sizes and the level of re-identification risk may differ from study to study. It is suggested that ARs should be more specific about the methodology for each study.

**Anonymisation of sensitive data**

Not all ARs provided information on the anonymisation of sensitive data. Sensitive data are not easily identified, may be atypical data points and hence may be missed by automated tools. The policy does not define what data should be considered sensitive. Fortunately, the General Data Protection Regulation (GDPR 2016/679) rectifies this omission and defines sensitive subject data as "race or ethnic origin, political opinions, religion or beliefs, trade-union membership, genetic data, data concerning health or sex life, or criminal convictions or related security measures."[3]

It is recommended that ARs should define what data are considered as sensitive and how these data are identified and anonymised.

**Pre- and post-anonymisation comparison**

The policy requires that there should be no difference in terms of content between primary use reports and anonymised or redacted reports. Not all ARs published provided information on meeting this requirement. The AR should describe any technical changes (formatting, pagination, hyperlinks) that may have occurred as a consequence of the anonymisation process.

**Data of deceased subjects**

The policy refers to data protection of "natural persons", thus excluding the deceased. Many trial subjects die during the course of a study as a consequence of underlying disease, yet their data are included in the report datasets.

Under the policy and the GDPR, these are no longer categorised as personal data. However, according to the Article 29 Data Protection Working Party, data on the deceased may be considered as personal information if they are linkable to living family members.[4]

The AR should specify how post-mortem data are dealt with during the anonymisation process.

## Conclusions

Anonymisation will rapidly evolve as technology continues to advance. The policy emphasises the importance of taking into account future developments when considering current anonymisation techniques. There is a trend towards the use of quantitative re-identification risk assessment and automated or artificial intelligence systems in anonymisation as the industry develops new tools and gains experience.

To the author's knowledge, this paper provides the first analysis of ARs that have been published on the EMA clinical website. The site was found to be a very useful and user-friendly resource for this type of research. A guide to

navigating the site is provided in an article beginning on p. 17.

## Acknowledgements

The author would like to thank Achim Schneider for his assistance in the analysis and quality check, and Louise Martinsson for her review.

## 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 is employed by a clinical research organisation and may have been involved in the writing of clinical reports in the EMA clinical website.

## References

1. European Medicines Agency. External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use (EMA Policy 0070). EMA/90915/2016. Version 1.0 (2 March 2016) Version 1.1 (16 December 2016), Version 1.2 (11 April 2017), Version 1.3 (20 September 2017). Available at: http://www.ema.europa.eu/docs/en_GB/documentvlibrary/Regulatory_and_procedural_guideline/2017/09/WC500235371.pdf.
2. Anonymisation Report. Procedure # EMEA/H/C/002345/II/0020. Version 29 August 2017. Published 04 October 2017 at https://clinicaldata. ema. europa.eu. (Accessed 30 December 2017.)
3. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN.
4. Article 29 Data Protection Working Party. Opinion 4/2007 on the concept of personal data. Technical Report 01248/07/EN WP 136, June 2007. Available at http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf.

## Author information

**Raquel Billiones** (Head, Medical & Regulatory Writing at Clinipace Worldwide) has been a regulatory medical writer since 2006, covering both pharmaceutical and medical device industries. Her core competencies include disclosure and data protection in clinical trial data reporting. She has served in various EMWA roles, including as EC member (2015–2017), associate editor for *Medical Writing* (2010–present), internship advocate, and workshop leader.

# The end of a successful pilot… GAPP retires after 6 years

The Global Alliance of Publication Professionals (GAPP) was set up in 2012 as a rapid response mechanism to counter misinformed or outdated articles about professional medical writing, publication planning and pharmaceutical industry sponsored publications. The original team incorporated individuals who served in leadership roles in the three main professional organisations – the American and European Medical Writers Associations (AMWA, EMWA) and the International Society for Medical Publication Professionals (ISMPP) – and who were located in different regions of the globe. Individuals have come and gone since the founding team, but we have always maintained that mix and geographic spread.

The conversation about industry funded research has been moving toward data disclosure and clinical trial transparency for some time. The volume of articles about industry sponsored medical publications has decreased, and though misinformed or poorly researched articles are still being published, they have been focussing less and less on professional medical writing support. We responded to only four articles in 2017.

So, after 6 years and ~50 responses, the GAPP feels the organisation has served its purpose and the professional organisations have evolved to a point that we can hand back responsibility for rebuttals and responses to the respective governing bodies. We are confident that the professional organisations will act in concert to develop timely and influential responses that will serve to educate our colleagues and our critics. We are therefore announcing the retirement of GAPP, effective as of the close of the upcoming ISMPP meeting on May 2, 2018.

The GAPP website and response archive will be maintained, and the contact@gappteam.org email address will still be monitored for any member of our profession to report an article they think justifies a response.

The retiring GAPP team would like to thank previous GAPP members as well as those who have referred articles to us over the years, and urges people to continue to be vigilant for inaccurate, misinformed, or outdated articles about our profession.

**Jackie Marchington,** Caudex, a McCann Health Company, Oxford, UK; ISMPP Advocacy & Outreach Committee co-Chair
**Cindy Hamilton,** USA; Past AMWA President and Founding Member of GAPP
**Art Gertel,** Principal, MedSciCom, LLC, Lebanon, USA; Past AMWA President and Founding Member of GAPP
**Serina Stretton,** Proscribe, part of the Envision Pharma Group, Macquarie Park, Australia
**Julia Donnelly,** Julia Donnelly Solutions, Derbyshire, UK; Past EMWA President.
http://gappteam.org
contact@gappteam.org