Anonymisation: A new challenge for medical writers

Montserrat Cuadrado Lafoz,[1] **Anna Ramírez-Soriano,**[1] and **Sarah Richardson**[2]

1 PPD, Barcelona, Spain

2 PPD, Cambridge, UK

Correspondence to:

Montserrat Cuadrado Lafoz
PPD
Torre Nozar, c/Titán, 15
28045 Madrid, Spain
+34 93 0111403
Montserrat.CuadradoLafoz@ppdi.com

Abstract

In its commitment to transparency, the EMA implemented Policy/0043 and Policy/0070 to make data accessible to all; however, this has given rise to the need for anonymisation of personal data in clinical reports. The analysis of the 64 submission packages containing anonymised data submitted to the EMA as of March 2018 shows that the most frequent technique to anonymise data is redaction. This is typically performed after reports are submitted to the competent authorities. The study report team, through a strong cross-functional strategy, can reduce the anonymisation required after submission of the document by proactively reducing the use of unnecessary identifiers in the initial report drafts. Therefore, the challenge for medical writers is to become involved in the anonymisation strategy and the creation of a data anonymisation plan for the clinical documents from their initial stages, focusing on the balance between scientific value and risk of re-identification.

Introduction

In its commitment to openness and transparency, the EMA initially implemented Policy/0043[1] followed by Policy/0070[2] on the publication of clinical data for medicinal products for human use. The primary objective of Policy/0070[2] is to make data accessible to all; however, the implementation guidance for these policies[3] has given rise to the need for **anonymisation** of **personal data** in clinical reports, which per Policy/0070 includes clinical overviews, clinical summaries, and clinical study reports including appendices 16.1.1 (Protocols and Protocol Amendments), 16.1.2 (Sample Case Report Form), and 16.1.9 (Documentation of Statistical Methods).

Personal data protection is a fundamental right in many countries. In the European Union, this right is protected by European legislation[4-7] and agency directives.[2,3,8] Policy/0070 is fully compliant with the applicable regulations (in particular Regulation 45/20016 and Directive 95/46/EC7). Applicants/Marketing Authorisation Holders are required to submit clinical reports that have been rendered anonymous, meaning that, data must be written in a form that does not identify individuals. The anonymisation strategy should represent the best balance between data utility (maximal retention of scientifically useful information) and an acceptably low risk of re-identification.

Clinical reports contain **direct identifiers** and indirect or **quasi-identifiers.**[9] Direct identifiers are elements that permit direct recognition or communication with the corresponding individuals, such as name, email, phone number, or subject identifier. Quasi-identifiers are variables representing an individual's background information that can indirectly identify that individual (e.g., geographical location, dates, or demographic data).

The purpose of this analysis was to determine the most frequently used anonymisation techniques for direct and quasi-identifiers. In addition, we consider how medical writers can positively impact the anonymisation process by initiating anonymisation at the time of writing clinical reports. The goal is to reduce required anonymisation efforts after publishing, thus aligning with the EMA requirement to publish clinical data without jeopardising personal data protection.

Methods

The EMA clinical data website (https://clinicaldata.ema.europa.eu/web/cdp/home) was accessed under the academic and other non-commercial research purposes 'terms of use'. An advanced search for clinical reports published between October 2016 (the first date that the anonymisation reports were available in the database) and March 29, 2018, was performed.

A total of 86 entries listed by product name were obtained. The search results were exported into an Excel file. If there was more than one entry for the same product name, only the entry with the submission package containing the highest number of documents (including clinical overviews, clinical summaries, clinical reports, and anonymisation reports) was selected for analysis – for example, there were two entries for Humira (product name), one including eight documents and one including 10 documents; only the entry that included 10 documents was selected for analysis. Each of the selected entries (77 in total) was accessed and the anonymisation report was downloaded and reviewed. To determine the individual anonymisation techniques and identifiers used in each submission package, direct identifiers, quasi-identifiers, and the techniques used for anonymising them were recorded in the Excel export file. To create summaries of the anonymisation techniques by submission package and by type of identifier, the anonymisation techniques identified in the anonymisation reports were classified into the following categories:

Most of the submission packages that were anonymised (57 [89%]) used redaction only as the anonymisation technique for direct identifiers and quasi-identifiers.

- **Redaction**: included the terms "redaction" and "masking"
- Preserving **pseudonymisation**: included preserving anonymisation only
- **Randomisation**: included use of random offset dates and use of random values within the study inclusion criteria
- **Generalisation**: included generalisation of age (in years) to 5-, 10-, and 20-year intervals and generalisation of medical history terms to high level term, high level group, or system organ class
- **Suppression**: included suppression and replacement with alternate text used in the anonymisation

The number and frequency of submission packages were calculated by anonymisation technique category (or combination of categories). For the submission packages that used a combination of categories, the number and frequency of packages by type of identifier and by anonymisation technique category were also calculated. The Excel file was also used to perform specific subanalyses by type of drug (orphan drugs, generic and biosimilar medicines) using the same approach.

A glossary of anonymisation-related terms used in this article is presented in Table 1. All terms included in the glossary are written in bold font on their first use in this article.

Results

As of March 2018, applicants/Marketing Authorisation Holders have published anonymised clinical reports for 77 medicines on the EMA website, including orphan drugs, generic and biosimilar medicines, as well as medicines for use in children.

Thirteen (17%) of the 77 submission packages were not anonymised. Twelve were not anonymised because only the clinical overview was published, and the other because no documents were included in the package. These were excluded from the analysis, leaving 64 submission packages in our analysis.

Table 1. Glossary of terms

Anonymisation	The process of rendering data into a form that does not identify individuals and where identification is not likely to take place.
Direct identifiers	Elements that permit direct recognition or communication with the corresponding individuals (e.g. name, email, phone number, or subject identifier).
Generalisation	Consists of generalising or 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 a week).
Personal data	Any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, particularly by reference to an identification number or to one or more factors specific to their physical, physiological, mental, economic, cultural or social identity (Article 2[a] of Regulation [EC] No 45/2001).
Pseudonymisation	Consists of replacing one attribute (typically a unique attribute) in a record with another. The natural person is still likely to be identified indirectly. Pseudonymisation reduces the risk of association of a dataset with the original identity of a data subject.
Quasi-identifiers	Variables representing an individual's background information that can indirectly identify individuals such as their date of birth, date of death, or date of clinic visit, residence postal code, sex and ethnicity. Quasi-identifiers also include demographics and socioeconomic information.
Randomisation	A family of techniques that alters the veracity of the data to remove the strong link between the data and the individual.
Redaction	Masking the data and text to be removed, often using a black box.
Suppression	The original value is replaced with alternate text. There are several ways that the replacement text can be selected, depending on the type of personal data and the readability in the anonymised reports.
Transformation	A process that reduces the risk of identifying a data subject by altering personally identifiable information in a dataset.

Table 2: Summary of the anonymisation techniques used for direct identifiers in the seven submission packages using a combination of techniques

Direct identifiers associated with trial individuals	**Preserving pseudonymisation N=7 n (%)**	**Suppression N=7 n (%)**	**Redaction N=7 n (%)**
Subject ID	7 (100.0)	–	
Screening number	2 (28.6)	–	2 (28.6)[a]
Accession number	2 (28.6)	–	1 (14.3)[a]
Manufacturer control number	2 (28.6)	–	1 (14.3)[a]
Patient ID	1 (14.3)	–	1 (14.3)[a]
Treatment ID	1 (14.3)	–	1 (14.3)[a]
Initials	1 (14.3)	–	1 (14.3)[a]
Sample ID	1 (14.3)	–	1 (14.3)[a]
Lot number	1 (14.3)	–	1 (14.3)[a]
Data clarification form ID	1 (14.3)	–	1 (14.3)[a]
Ticket number	1 (14.3)	–	1 (14.3)[a]
Sample reference number	1 (14.3)	–	–
Barcode	1 (14.3)	–	–
Custom ID	1 (14.3)	–	–
Kit number	1 (14.3)	–	–
Pharmacogenomic ID	1 (14.3)	–	–
Photos of trial individuals	1 (14.3)	–	–
Experiment number[b]	–	1 (14.3)	–
Oceans ID[c]	–	1 (14.3)	–
Direct identifiers associated with staff information			
Non-investigator staff names[d]	–	7 (100.0)	–
Sponsor and staff contact details[d]	–	7 (100.0)	–
Site ID[e]	–	5 (71.4)	–
Contract research organisation[f]	–	–	–
Signature	–	–	7 (100.0)

Abbreviations: ID, identifier.

a Redaction was used when the direct identifier was embedded in an image.

b Direct identifier was assigned to a subject sample in testing for the investigational product binding in an investigational product antibody assay.

c Oceans ID was a direct identifier assigned to a subject specifically in the event of development of progression of progressive multifocal leukoencephalopathy due to the investigational product.

d Suppression was used unless the information was associated with a subject ID, in which case this information was treated as a quasi-identifier associated with the subject ID.

e Site ID was considered a direct identifier when it was in the presence of staff but not in the presence of a trial individual. In one of the packages, the site ID was associated with a subject ID; the information was treated as a quasi-identifier associated with the subject ID.

f Suppression was used where the contract research organisation was named in the context of a staff member.

Anonymisation techniques for direct identifiers

Most of the submission packages that were anonymised (57 [89%]) used redaction only as the anonymisation technique for direct identifiers. The remaining seven (11%) submission packages used combinations of redaction and preserving pseudonymisation (six packages [9%]); or redaction, preserving pseudonymisation, and suppression (one package [2%]) for direct identifiers associated with trial individuals.

All seven submission packages used the combination of redaction and suppression for direct identifiers associated with staff information. The sponsor name was always retained, and the coordinating or site investigator names were retained unless they were directly associated with a subject.

Table 2 shows the anonymisation techniques used for each identifier for submission packages that used a combination of techniques.

Anonymisation techniques for quasi-identifiers

Most of the submission packages that were anonymised (57 [89%]) used redaction only as the anonymisation technique for quasi-identifiers. The remaining submission packages used a combination of redaction, suppression, generalisation, and randomisation (six packages [9%]); or redaction and randomisation (one package [2%]) for quasi-identifiers associated with trial individuals.

Table 3 shows the anonymisation techniques used for each quasi-identifier for submission packages that used a combination of techniques.

> The study report team, led by a medical writer with a solid knowledge of EMA Policies 00431 and 00702 and the policies' implications for data protection, can reduce the anonymisation required after submission of the document by proactively reducing the use of unnecessary identifiers in the initial report drafts.

Table 3: Summary of the anonymisation techniques used for quasi-identifiers in the seven submission packages using a combination of techniques

	Generalisation	**Suppression**	**Randomisation**	**Redaction**[a]
	N=7	**N=7**	**N=7**	**N=7**
Quasi-identifiers	**n (%)**	**n (%)**	**n (%)**	**n (%)**
Age	6 (85.7)	3 (42.9)	5 (71.4)[b]	–
Race	–	4 (57.1)	–	–
Ethnicity	–	2 (28.6)	–	–
Height	–	2 (28.6)	–	–
Weight	–	2 (28.6)	–	–
Body mass index	–	2 (28.6)	–	–
Lean body mass	–	2 (28.6)	–	–
Body surface area	–	2 (28.6)	–	–
Waist	–	1 (14.3)	–	–
Drinking habits	–	3(42.9)	–	–
Smoking habits	–	3(42.9)	–	–
Family circumstances	–	1 (14.3)	–	–
Social circumstances	–	1 (14.3)	–	–
Medical history	6 (85.7)	6 (85.7)	–	–
Family medical history	–	2 (28.6)	–	–
Psychiatric hospitalisation	–	1 (14.3)	–	–
Dates	3 (42.9)	5 (71.4)	6 (85.7)[c]	–
Site ID	–	1 (14.3)	4 (57.1)[d]	–
Name	–	1 (14.3)	–	–
Staff name	–	1 (14.3)	–	–
Country	–	4 (57.1)	–	–
State	–	2 (28.6)	–	–
City	–	1 (14.3)	–	–
Region	–	1 (14.3)	–	–
Company/contract research organisation addresses	–	–	–	1 (14.3)
Organisation	–	2 (28.6)	–	–

Abbreviations: ID, identifier

[a] In three packages quasi-identifiers were redacted when these were embedded in images and could not be otherwise transformed. In three packages narratives were redacted; in three packages, listings were also redacted. The specific quasi-identifiers redacted in these packages were not specified.

[b] Age was supressed and replaced with a random value within the age range of the study population.

[c] The clinical dates were 'PhUSE offset' in five of the six packages; in one package calendar dates were adjusted based on an offset date.

[d] Site ID was supressed and replaced with a site ID chosen at random from the study.

The disadvantage of using redaction as an anonymisation technique, as opposed to techniques such as generalisation or randomisation, is that clinically relevant data that may be important in the context of the disease is lost in the redaction process.

Generalisation was used to transform age and medical history terms. The original age was replaced with a random age selected within an interval (5-, 10-, or 20-year intervals). The original medical term was replaced with a string of text corresponding to the high-level term, high-level group, or system organ class. The level of generalisation was dependent on the risk of re-identification for the individual subject (e.g., a subject could have his/her age generalised to a 10-year interval while another subject with lower risk could have his/her age generalised to a 5-year interval).

Randomisation was commonly used for dates and ages. Dates were replaced with a new date generated using a random offset for each individual and this offset was applied to all dates in the study for that individual. The most common algorithm used to offset dates was the PhUSE offset[10] (six of the seven submission packages that applied randomisation used this algorithm). Age was replaced with a random value within the age range of the study population.

Suppression was used for other quasi-identifiers such as race, ethnicity, weight, height, and body mass index. The alternate text used to replace the original value was dependent on the type of personal data and the readability in the anonymised reports; e.g., in some instances the alternate text was longer than the original text and tables could be difficult to read.

According to the anonymisation reports analysed, more than one anonymisation technique could be used for a given quasi-identifier; e.g., generalisation to a specific year interval or replacement with a random value within the inclusion criteria could be used for age, depending on the risk of re-identification.

Anonymisation techniques by type of drug

Of the 77 submission packages identified, 14 corresponded to orphan drugs, 14 to generics, and two to biosimilar medicines. The remaining 47 submission packages were not classified in any of these categories.

Of the 12 submission packages that were not anonymised because only the clinical overview was published, seven corresponded to generics. The remaining five submission packages were not classified as orphan drugs or biosimilar medicines. The submission package for which no anonymisation technique was used (because no documents were included in the package) was for an orphan drug. These submission packages were excluded from the analysis.

Anonymisation techniques for direct identifiers by type of drug

All of the generic and biosimilar medicines submission packages that were anonymised used only redaction as the anonymisation technique for direct identifiers. A total of 10 (77%) of the 13 orphan drug submission packages that were anonymised used only redaction as the anonymisation technique for direct identifiers.

Three (23%) of the orphan drug submission packages that were anonymised used redaction together with other techniques; these equate to almost half (43%) of the seven submission packages overall that used redaction together with another anonymisation techniques.

Of the three orphan drug submission packages that were anonymised using redaction together with other techniques, two used a combination of redaction and preserving pseudonymisation; and one used redaction, preserving pseudonymisation, and suppression for direct identifiers associated with trial individuals. All three anonymised orphan drug submission packages used redaction and suppression for direct identifiers corresponding to staff information.

Anonymisation techniques for quasi-identifiers by type of drug

All of the generic and biosimilar packages used only redaction as the anonymisation technique for quasi-identifiers, except for one generic package that used redaction in combination with randomisation (calendar dates were adjusted based on an offset date).

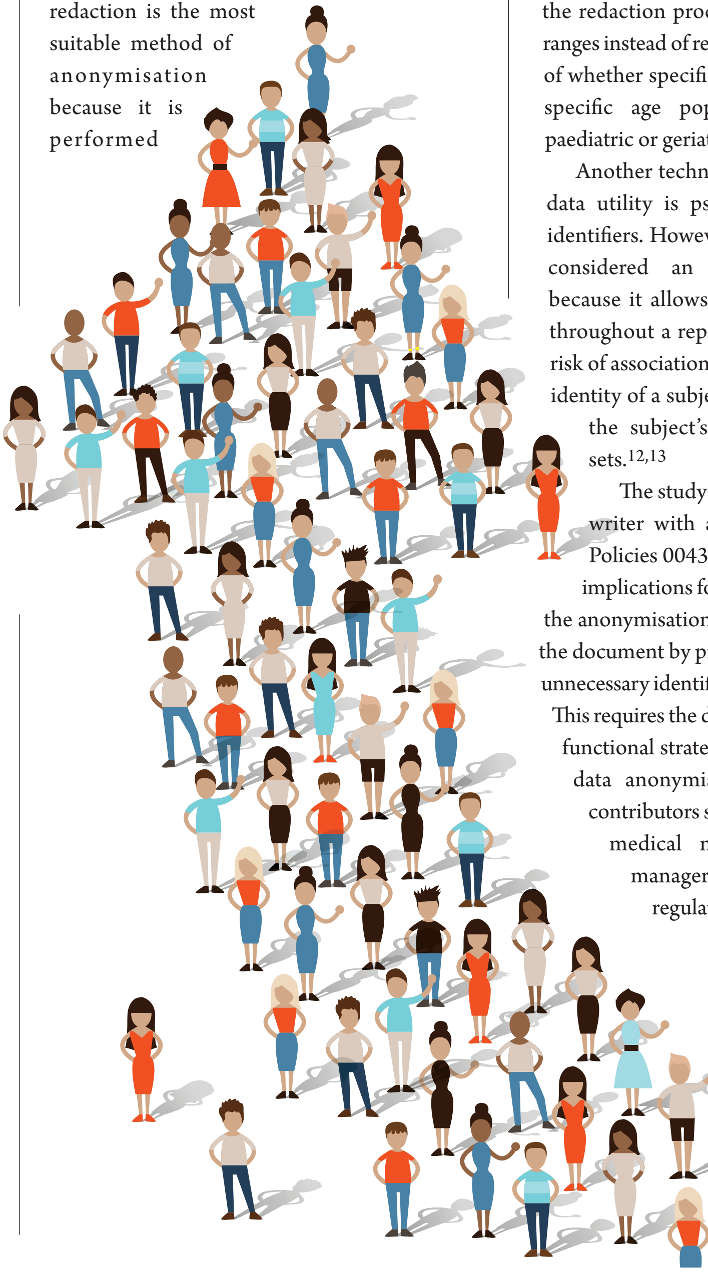
A total of 10 of the 13 (77%) orphan drug packages used only redaction as the anonymisation technique for direct identifiers.

Three (21%) orphan drug submission packages used redaction together with other anonymisation techniques (a combination of redaction, suppression, generalisation, and randomisation). These three orphan drug submission packages equate to almost half (43%) of the seven submission packages overall that used redaction together with other anonymisation techniques.

Discussion

The results of this analysis show that the most frequently used approach to anonymise data in reports submitted to the EMA is redaction of identifiable personal data that has been included in the original document. The use of other anonymisation techniques, such as **transformation** or generalisation of identifiers, is generally limited, and most frequent in orphan drug reports.

In addition, our analysis shows that anonymisation is typically performed after reports are submitted to the competent authorities. Therefore, redaction is the most suitable method of anonymisation because it is performed retrospectively. These findings are supported by Kumar and Sareen[11], who suggested several reasons for using redaction only: (1) most of the documents are anonymised retrospectively, (2) most of the automated anonymisation tools are proficient only in performing redaction, and (3) application of other techniques in addition to redaction makes the process more time consuming.



The disadvantage of using redaction as an anonymisation technique, as opposed to techniques such as generalisation or randomisation, is that clinically relevant data that may be important in the context of the disease is lost in the redaction process. For example, using age ranges instead of redaction allows determination of whether specific findings are only related to specific age population groups (such as paediatric or geriatric subjects).

Another technique that is used to preserve data utility is pseudonymisation of subject identifiers. However, pseudonymisation is not considered an anonymisation technique because it allows for a subject to be tracked throughout a report. Although it reduces the risk of association of a dataset with the original identity of a subject, it is still possible to track the subject's data across different data sets.[12,13]

The study report team, led by a medical writer with a solid knowledge of EMA Policies 0043[1] and 0070[2] and the policies' implications for data protection, can reduce the anonymisation required after submission of the document by proactively reducing the use of unnecessary identifiers in the initial report drafts. This requires the development of a strong cross-functional strategy on data anonymisation (a data anonymisation plan) involving key contributors such as the medical writer, the medical monitor, the clinical data manager, the biostatistician, and the regulatory affairs representative.

The data anonymisation plan may involve several different strategies to reduce the risk of re-identification while maintaining data utility. These may include the use of age ranges and pseudonymising as previously discussed. Other strategies are to avoid the use of subject identifiers in the body of the report; generalise from country to region; avoid the use of gender-related words; and present relative days (e.g., day since drug administration) rather than the actual dates.

Therefore, the challenge for medical writers is to become involved in the anonymisation strategy and the creation of a data anonymisation plan for the clinical documents from their initial stages, focusing on the balance between scientific value and risk of re-identification, especially for studies involving small populations and on rare diseases.

Acknowledgements

We thank Tracy Farrow and Amy Dorman of PPD for their support in reviewing and editing this article.

Conflicts of interest

The authors are employed by PPD.

References

1. EMA. EMA policy on access to documents (related to medicinal products for human and veterinary use). Policy/0043 (EMA/110196/2006). 2010 [cited 2018 Jun 13]. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2010/11/WC500099473.pdf.
2. EMA. EMA policy on publication of clinical data for medicinal products for human use. Policy/0070 (EMA/240810/2013). 2014 [cited 2018 Jun 13]. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf.
3. EMA. External guidance on the implementation of the EMA policy on the publication of clinical data for medicinal products for human use (EMA/90915/2016). 2017 [cited 2018 Jun 13]. http://www.ema.europa.eu/docs/envGB/document_library/Regulatory_and_procedural_guideline/2017/09/WC500235371.pdf.
4. Code of Federal Regulations. Title 45: Public Welfare, Subtitle A §164.514. [cited 2018 Jun 13]. https://www.gpo.gov/fdsys/pkg/CFR-2002-title45-vol1/pdf/CFR-2002-title45-vol1-sec164-514.pdf.
5. Health Insurance Portability and Accountability Act Privacy Rule. [cited 2018 Jun 13]. https://www.hhs.gov/hipaa/for-professionals/privacy/index.html.
6. Council Regulation (EC) 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the community institutions and bodies and on the free movement of such data. 2008 [cited 2018 Jun 13]. OJ L193/7. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32001R0045&from=EN.
7. Council Directive (EC) 95/46 of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. 1995 [cited 2018 Jun 13]. OJ L281. https://ec.europa.eu/health/sites/health/files/files/eudralex/vol-1/reg_2014_536/reg_2014_536 _en.pdf.
8. Regulation (EU) 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC. 2014 [cited 2018 Jun 13]. OJ L158/1. https://ec.europa.eu/health/sites/health/files/files/eudralex/vol-1/reg_2014_536/reg_2014_536_en.pdf.
9. Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. BMJ. 2010;340:c181.
10. Ferran JM, El Emam K, Nolan S, Grimm B, De Donder N. PhUSE De-Identification Working Group: Providing De-Identification Standards to CDISC Data Models. Presented at the PhUSE Annual Conference, Vienna, October 11-14, 2015 (Paper DH01). Available from: https://www.phusewiki.org/docs/Conference%202015%20DH%20Papers/DH01.pdf.
11. Kumar P, Sareen R. Evaluation of Re-identification Risk for Anonymized Clinical Documents. Presented at the PhUSE Annual Conference, Edinburgh, October 8-11, 2017 (Paper RG02). Available from: http://www.phusewiki.org/docs/Conference%202017%20RG%20Papers/RG02.pdf.
12. Data Protection Working Party. Opinion 05/2014 on Anonymisation Techniques. Article 29. WP 216. Available from: http://www.dataprotection.ro/servlet/ViewDocument?id=1085.
13. Data Protection Working Party. Opinion 4/2007 on the Concept of Personal Data. Article 29. WP 136. Available from: https://www.clinicalstudydatarequest.com/Documents/Privacy-European-guidance.pdf.

Author information

Montserrat Cuadrado Lafoz, PhD, is a senior medical writer within the PPD Global Medical Writing Group. Montse has 7 years of experience writing regulatory documents.

Anna Ramírez-Soriano, PhD, is a program manager within the PPD Global Medical Writing Group. Anna has 9 years of experience writing regulatory documents, and manages a diverse portfolio of work for a MW functional service partnership. She also leads a module in the Pharmaceutical Industry Master at University of Barcelona.

Sarah Richardson is a Principal Medical Writer within the PPD Global Medial Writing Group. Sarah has 7 years of experience writing regulatory documents, and is compound lead for 2 programmes within a medical writing functional service partnership. She manages the writing team and oversees all deliverables for these programmes.