Retrospective vs. Proactive Anonymization of Narratives

Project Co-leads – Rashmi Dodia, MS, RAC and Gregory Campbell, BS

Executive Summary

Ever since policy 0070 was first introduced in 2016, there have been continuing concerns and challenges among sponsors with respect to its implementation across clinical study reports (CSRs), in particular the patient narratives section. As per the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) E3 (Section 12.3.2), a patient narrative should describe **deaths, serious adverse events (SAEs),** or **adverse events (AEs)** of special clinical interest and those leading to permanent discontinuation from a clinical trial and should include key information like patient identifiers, age, sex and race of the patient among other indicators like their disease/medical history, concomitant medications, current treatment outcomes and test results. Needless to say, with the increased amount of patient information, there is also an increased level of effort required to produce anonymized narratives that strike the right balance between patient confidentiality protection and data utility.

The analysis published by Khaled El Emam in April 2017 pointed towards a majority of sponsors opting to redact narratives entirely from their marketing authorization applications.[1] This trend seems to have continued on since then. However, it is important to acknowledge that narratives offer further context towards understanding an event by providing key information like the nature of the event in verbatim terms, and also the timing of occurrence. Narratives also provide information on what tests were run on a particular patient (even if the actual test outcomes are anonymized), which could be another indicator of certain types of conditions that a researcher could potentially benefit from. Many researchers agree that narratives are essential in understanding related higher level concepts for events, which in turn plays an important role in any potential downstream meta-analysis.[2] Since policy 0070 is geared towards increasing data utility, sponsors should start exploring ways of anonymizing narratives instead of completely redacting them to preserve the essence of the policy by allowing academics and researchers to re-assess clinical data.

This White Paper focuses on two approaches to produce anonymized narratives – retrospective and proactive. The retrospective section sheds light on the challenges faced with qualitative methods like redaction and what impact it has on data utility. Given the limiting nature of retrospective anonymization with regards to data utility, there is clearly a need for modern solutions and enhanced skills to be developed in order to meet Policy 0070 requirements. This paper factors in the possibility of integrating a tool or software solution that supports retrospective anonymization and defines business requirements that will help sponsors make an informed decision on automating some of the anonymization efforts. The second part of this White Paper offers a perspective on proactive anonymization and how to operationalize it.

Introduction

EMA Policy 0070 states – ***"Case narratives should not be removed nor redacted in full regardless of their location in the clinical reports (body of the report or listings). They should be, instead, anonymised. Regardless of the anonymisation technique used by the applicant/MAH, EMA cannot accept the redaction of the entire case narrative by default (as a rule). If, exceptionally, the entire case narrative needs to be redacted to ensure anonymization, i.e. all identifiers (direct and indirect) need to be redacted; it has to be clearly justified in the anonymization report."***

ICH E3 requires that the overall adverse event experience in the study should be described in a brief narrative, supported by the following more detailed tabulations and analyses. Such brief narratives describing deaths, other serious and significant adverse events be placed either in the text of the CSR or in section 14.3.3, depending on their number.[3]

Retrospective Anonymization of Narratives – Considerations

Narratives included in CSRs contain large amounts of personal participant information that includes direct identifiers in association with many indirect identifiers that could be used to re-identify an individual. Retrospective anonymization of unstructured data is challenging, as protected personal data (PPD)/personally identifiable information (PII) can appear virtually anywhere in a clinical narrative. Protecting patient confidentiality is a requirement, an expectation's that it should not be overlooked or understated, especially with the increasing technological advancements in the area of transparency. With greater transparency and increasing amount of access, it has become even more critical to establish the right controls for the protection of personal data. For clinical data to be considered anonymized, there are certain data elements that will have to be removed, depending on the technique one utilizes for anonymization.

For legacy studies requiring retrospective processing, narratives are commonly redacted (using Adobe Acrobat) either partially or in full. The redaction tool within Adobe Acrobat allows reviewers to clearly identify the text that is marked for redaction, apply necessary overlay codes (mandated by Policy 0070) and be able to completely remove/block the text that is proposed for redaction by applying the redaction marks. Some of the more advanced techniques require conversion of PDF files to other formats for further processing, including full data anonymization, which can have huge implications on cost and content control. A complete manual review for efficient redaction of narratives could turn out to be a cumbersome process and may produce inconsistent outputs depending on subjective interpretation of rules used to redact patient level information. In order to overcome some of these inconsistencies and deliver, sponsors should develop a defined set of rules to cater to different study types and ensure redactors have a deep understanding of these rules via training, work practices and guidance documents. Since manual processes are prone to human error, a detailed quality control process at the end should also be implemented to ensure consistent application of rules across all narratives. Stringent training exercises and thorough QC checkpoints will enable the highest level of accuracy. Using partially automated techniques with pattern/string matching and batch searching for keywords via simple Adobe plugins will be helpful in reducing the manual review burden to some extent. At present, there are no befitting tools to achieve automated retrospective anonymization of unstructured data like narratives.

When considering anonymization of narratives, depending on the study sample size and disease category, some or all of the following data elements need to be masked, to ensure they are sufficiently anonymized. The list below includes HIPAA safe harbor elements (these only serve as a reference point for systematic anonymization and should not be considered as the gold standard).[4] The determination as to what elements need to be masked and the level of masking required for each, solely depends on a sponsor's approach to anonymization and data utility considerations. Furthermore, due consideration will have to be given to patients' informed consent when designing a study-driven approach. Most sponsors are currently employing a "qualitative" approach towards anonymizing their clinical study documents. This approach assesses relative risk of re-identification based on the level of information provided by direct and indirect identifiers associated with patients. These identifiers are described in further detail below:

- Patient/Subject IDs, CIOMS/Medwatch numbers, other identifying numbers
  - Direct identifiers; high risk element
- Geographic locations, race and ethnic characteristics
  - Indirect identifiers; high risk element
  - Could have an undesirable impact on data utility if the efficacy/safety analyses of a study are based on geographic locations
  - Linkability with other identifiers could pose an issue
- Age
  - Indirect identifier; high risk element
  - Could have a substantial impact on data utility if the efficacy/safety analyses of a study are based on age groups
  - Linkability with other identifiers could pose an issue

- All dates related to an individual (birth, death, treatment-related)
  - Indirect identifiers; high risk element (specially birth and death dates)
  - Anonymization strategy for treatment dates could have a direct impact on data utility
  - Linkability with other identifiers could pose an issue
- Medical history/concomitant medications, etc.
  - Indirect identifiers; moderate risk element (could be considered high risk particularly with trials of highly targeted therapies for small patient populations)
  - Linkability with other identifiers could pose an issue
- Physical characteristics such as height, weight, body mass index (BMI)
  - Indirect identifiers; low risk element
  - Linkability with other identifiers could pose an issue

In certain instances, special events could be particularly identifying too. For example, events that could be reported over local news channels – such as suicides, homicides, accidents, etc. These may require special attention.

Moreover, a different rule set will have to be developed by sponsors to cater to different types of studies, based upon some of the study characteristics listed below. This list is not exhaustive and only includes potential factors to be taken into consideration:

- Study sample size
- Disease prevalence (rare/orphan diseases)
- Studies including sensitive data
- Stigmatizing disease studies
- Number of study sites
- Number of patients per site
- Sites per country
- Specific geographical locations with small populations

A few possible scenarios are listed below in Table 1. Based on a paper published by TransCelerate on "Protection of Personal Data in Clinical Documents", Table 1 takes into consideration various types of studies and provides an anonymization approach that outlines which identifiers could be masked/redacted in narratives to avoid the risk of re-identification (Column B).[5] Based on the redactions proposed in column B, column C lists identifiers that are left unredacted in each case, and the impact on data utility and re-identification.[6]

A: Scenario	B: Redaction Approach	C: Identifiers left unredacted and potential impact on data utility
1) Large Multicenter Study with more than 100 participants enrolled globally	Mask/Redact all direct identifiers (subject IDs, CIOMS/Medwatch or other patient identifying numbers) and the following limited indirect identifiers: • Age • Race/geographic locations (appearing in association with individual patients) • Dates - birth/death dates and treatment related dates (per HIPAA, for patients below 89 years of age, all elements of dates, except, year should be redacted). For patients aged 90 and above, year should be redacted too, along with other elements of dates. However, to ease the burden on manual review, a general approach could be adopted to redact patient related dates in their entirety for all age groups • Medical history, prior treatments, concomitant illnesses and medications, etc. • Free descriptive text in the CSR (brief narratives/1-2 line listings (containing unique information on patients, leading to their easy identification) • Specific incidents/accidents that could be known via local news, social media, etc.	By adopting this approach, the identifiers that are left unredacted include: • Sex • Adverse event descriptions and associated treatments • Treatment outcomes and test results • Physical characteristics such as height, weight and BMI Impact on data utility: The text that remains after redaction of identifiers proposed in column B will preserve association between adverse events, associated treatments and outcomes/test results as they occur in either males or females, thus preserving a moderate level of safety data utility. However, since linkability with the patient's age, geographic location and medical history will not be preserved, it will negatively impact data utility. The redaction of patient IDs and dates will hamper the understanding of treatments performed, and their outcomes in an individual patient, throughout the clinical study, and even more so, in case of trials that are conducted for seasonal diseases. However, if patient visits are tied to the treatment administration, not all is lost.

A: Scenario	B: Redaction Approach	C: Identifiers left unredacted and potential impact on data utility
2) Smaller studies (mainly owing to disease prevalence) with fewer than 100 participants or single sites	Mask all direct and indirect identifiers listed above along with: ● Height, weight, BMI ● Treatment outcomes and test results	By adopting this approach, the identifiers that are left unredacted include: ● Sex ● Adverse event descriptions and associated treatments Impact on data utility: This approach does impact data utility to a large extent. With the redaction of identifiers mentioned in column B, the remaining data would be highly fragmented and missing vital pieces of information that could aid in the understanding of study related procedures and outcomes. However, in studies with fewer than 100 participants and/or single sites, the presentation of extensive patient specific information in narratives does warrant conservative redactions nonetheless, in order to protect patient privacy.
3) Study populations selected on the basis of, or with a notably high prevalence of, any of the following conditions: psychiatric disorders, reproductive disorders, sexually transmitted diseases, drug or alcohol abuse, pregnancy,	Remove narratives entirely	Given the sensitive nature of the study and the population enrolled in this category, the safest approach would be to remove narratives entirely in order to avoid linkability of any re-identifying information for specific patients.

A: Scenario	B: Redaction Approach	C: Identifiers left unredacted and potential impact on data utility
congenital abnormalities, special populations, rare diseases, etc.		

Even when defining rules for redactions based on study categories, a certain level of flexibility will have to be allowed in order to better protect patient anonymity. Some scenarios where deviations would apply are outlined below:

1) When redacting ages for patients that are 90 years old and above, it is important to be mindful of the total population that is in that age range. If only a few patients are above 89 years of age, adapting the threshold to a lower number would be a better approach to take to maintain anonymity in the elderly population age group (eg, redacting all ages 75 and above)
2) If a sponsor decides to leave the year in when redacting dates related to medical history in a large population study, they will still have to carefully take a closer look at the dates of events that are left in and ensure that they do not help identify the patient as very young or very old (eg, "Patient XYZ had a tonsillectomy in 1985" leads up to the patients age to at least ~40 years).

Redacted reports will probably be less valuable for research compared to other sources such, as anonymized datasets. Even though the same data elements are likely removed/manipulated in the datasets as in the documents, it is the structured nature of the datasets (and anonymized patient IDs linking throughout) that provide greater clinical utility. For example, with important data elements removed or anonymized (such as dates, location, medical history, etc.), it is difficult to analyze temporal associations between disease and treatment occurrences, or gauge the effect of diseases or findings on people of different race, ethnicities, etc.[6] Moreover, redacted patient IDs also prevent linkage between events for the same patient. In the absence of dates, it is difficult to understand the correct sequence of events and their relationship with treatments. So, such reports do not provide a complete picture of treatments and related efficacy, and safety outcomes of any one patient to researchers, preventing their full understanding at the patient level. It could also have an impact on the interpretation of overall study results. A few anonymization systems attempt to retain some temporal data by using 'date altering' methods, thereby preserving the intervals between dates. Such methods are difficult to implement retrospectively in clinical documents and would require regeneration of narratives from anonymized data. Other data potentially valuable to researchers, such as patient occupations and ages of elderly patients older than 89, are lost through the anonymization process and not present in anonymized reports. In order to achieve the right balance between data utility and risk of re-identification in a retrospective setting, a significant amount of time, analyses and resources are required. Dorr et al. have evaluated the time cost to manually anonymize narrative text notes (average of 87.2 ± 61 seconds per note), and concluded that it was time-consuming and difficult to exclude all PPD/PII required by HIPAA.[9]

Several anonymization systems have been introduced that attempt to automatically identify PPD/PII in unstructured data, such as free-text medical records, narratives, etc. But a gold standard has not been established yet and there could be implementation issues that may take a few years to resolve.[7,8] Some of these issues could be misclassification of clinical data as PPD/PII and removal of clinically relevant information during the anonymization process, use of the system across varied document types to establish a generalized approach, etc.[9] So, even though Policy 0070, guidance advises for anonymization of PPD in narratives rather than their complete removal, a mixed approach has been observed so far among the marketing authorization applications (MAAs) posted on the EMA clinical website, wherein certain sponsors are opting to remove narratives entirely before they submit their redaction packages to the EMA. With technological advancement, an increasing number of sponsors employing advanced anonymization techniques and issuance of updated guidance, the EMA may not be as accepting of the completely redacted narratives.

As expressed above in Table 1, sponsors need to have a defined strategy that could be applied to various types/categories of studies consistently. Having a standardized method to redact data can produce a scalable solution that can be adapted to meet sponsor requirements. The method can be tailored to meet Policy 0070 disclosure requirements and also balance data utility of redacted narratives with sponsor's responsibility to protect patient information.

A number of easily available software's, open platform packages, and sophisticated tools can be evaluated by sponsors to meet their needs. There are many simple Adobe plug-ins that offer the ability to create/import keyword dictionaries in the form of text files and allow batch processing of PDF documents, offering a much better solution to the limited search and processing features available in Adobe alone. Sponsors could also implement open-source script packages or anonymization software packages that include codes, and dictionaries for automated location, and removal of PPD/PII in free text. Alternatively, several advanced tools have been launched recently to produce redacted/anonymized documents, and automated risk assessment reports, specifically for Policy 0070. These tools and technologies are in their evolutionary stages and their full potential may be realized only after a few years of industry exposure.

Selection of one or a combination of these tools is solely dependent on the sponsor's business needs. Several factors have to be taken into consideration when evaluating these tools for implementation into retrospective anonymization processes. Some of these factors are described below:

1) Technical Features :

- **Interface**: The tool should be simple to understand and implemented with a user-friendly interface
- **Valid Technique**: The tool should also be validated against the requirements defined in Policy 0070 and sponsor SOPs, to produce accurate outputs every time. It should also have the ability to apply full-proof anonymization/redaction every time, where the original text cannot be recalled by data attack using advance techniques.
- **Input/Output Formats and Source Files**: Input data formats and output data requirements will determine a tool's feasibility with internal sponsor processes as it may get difficult to convert final published reports from MAAs to a different format accepted by the tool and then reformat as a PDF the tool-generated outputs for submission to the EMA. It would be much easier if the tool accepts published reports as input and generates anonymized PDFs at the end. It should also be able to produce consistent outputs with or without the use of individual patient datasets (IPD). Additionally, it should also be able to produce an intermediate version before the generation of a final anonymized output for allowing EMA review of the sponsor defined strategy during the proposal consultation phase of the Policy. If the tool is producing anonymized outputs, it should be able to handle string lengths so as not to allow interpretation of source values from the number of characters being masked by the anonymized term. And if the tool is producing redacted outputs, it should be able to overlay redaction marks with overlay text and color-code them, per EMA requirements.
- **Quality Control**: It should allow the ability to manually edit final outputs, if discrepancies are noted during quality control. It should also produce a table containing all anonymized text replacements made in a report along with their original text and page number, so as to aid the final review process.
- **Smart Processing**: The tool should be able to store anonymization keys in a secure fashion, so as to allow the use of similar anonymized patient ID and other identifier when an extension study of a previously anonymized study or a line extension of a previously anonymized product is used an input. It should also have the ability to store inputs as general dictionaries and create product-specific libraries for PPD/PII that could be used for future studies. Capacity to process natural language and conduct concept-based searches to identify certain disease groups for medical history, etc. is also required in the tool.
- **Anonymization Strategy and Reporting**: It is important to ensure the tool is adaptable to apply a sponsor-defined anonymization approach in a consistent manner, in the unstructured environment of narratives; especially when the risk levels and strategy are driven by study size and nature of the trial. It should also be customizable, so workflow changes could be

implemented based on guidance updates and industry standard evolution. It should also have the ability to report any deviations from the strategy in case of outliers. Additionally, it should be configurable to industry accepted risk levels of re-identification as the EMA will not accept submissions that do not meet the guidance-defined risk threshold. Finally, with the evolution of the global transparency landscape, it is important that the tool is customizable to implement country-specific requirements to a report with minimal amount of manipulation.

- **Document Management**: The tool should have the ability to integrate major DMS systems like SharePoint, and directly download sources and upload outputs on it, in order to reduce the manual oversight required for each dossier.

2) Resources: Based on the functional capabilities of the tool, the resources required per study to complete tasks such as initial requirements gathering and input of metadata in the tool, QC of the output, etc. will also have to be factored in. Ideally, the tool should be as resource neutral as possible with requirements for a dossier being defined and entered at the onset and only fine adjustments needed later on to obtain study-specific results. The cost of training resources and maintaining study level oversight may also differ for each tool and should be taken into account by the sponsor.

3) Licensing and Maintenance Costs: Another important factor to consider, is the initial setup phase including the cost of licensing or buying the tool, based on sponsor requirements and internal policies. Regular software updates, periodic outages for maintenance, license renewals will also have to be discussed in detail. A cost analysis of all available softwares and tools comparing overall cost per study, including resources, training, etc. will also have to be conducted to reach a final decision.

Policy 0070 guidance also emphasizes the need for sponsors to continuously monitor the development of technologies in this area in order to assess novel risks of re-identification for any future clinical reports published, and to also identify the most suitable technique (or a combination of techniques) to establish an adequate anonymization process for clinical documents. The future evolution of anonymization tools should focus on preserving maximum data utility, while ensuring adequate anonymization, in order to stay compliant with Policy 0070 requirements.[10,11] Even though the simplest method of masking is also the most prevalent one, efforts could be made to explore other techniques such as randomization and generalization in the context of anonymizing narratives. Certain trends for future exploration include:

1) Use automated systems to generate narratives wherein all patient-specific information (such as demographics, medical history, etc.) is provided upfront (preferably in a tabular format). It would be easier to anonymize using such systems than with redaction. The narrative part of the

description would then only focus on treatments and associated analyses and outcomes (without alluding to patient specifics). This method will aid in removing the strong link between the data and the individual.

Moreover, a number of natural language processing (NLP) platforms are being created and made available publicly to ease in the identification of unstructured PPD/PII. A combination of such tool with other automated software's could be helpful.

2) Using the same set of study-specific requirements to anonymize both datasets and narratives in parallel. This would ensure consistency of outputs and reduce subjective differences that are often seen when using techniques like redactions.

3) Use of anonymization techniques for generating anonymized datasets and generating another copy of the CSR using anonymized datasets (proactive anonymization). This technique is described below in further detail.

Another concept worth exploring is the idea of controlled release. Evaluating the target audience for narratives closely and providing controlled access to such individuals could minimize the chances of re-identification. Retrospectively redacted narratives combined with tabulated sections of the CSRs, could provide a near complete picture to researchers/academics that are interested in the onward value of clinical research. Part II of the policy could open doors for access to anonymized IPD for studies submitted as part of an MAA, even though it is already voluntarily done by some sponsors as part of their data-sharing initiatives. Moreover, it has been argued by Ferris and others, that to protect PPD/PII, a hybrid system approach utilizing anonymization and role-based access for IRB-approved researchers may be preferable to offer flexible control of PPD/PII, while meeting the needs of biomedical researchers.

Proactive Anonymization of Narratives:

It is very important to shift focus to anonymizing patient-level data at an early stage to ensure that downstream uses of those data are secure, including use in clinical reports, ICH section 14 tables and narratives. This method would also overcome the issue of misinterpretation of results that are highlighted in the section above on retrospective anonymization (a consequence of ad hoc masking of the data).

Anonymization techniques enable publication of detailed information, which permits ad hoc queries and analyses, while minimizing the exposure of private or sensitive information in the data against a variety of attacks. Several approaches currently exist and are used by sponsors to effectively anonymize data so that detailed results can be published and shared with others. The most attractive feature is that they preserve the actual characteristics and statistical properties of original data. Hence, they provide a wide opportunity to share data for research purposes while maintaining utility and anonymity. Since most sponsors already have internal anonymization standards established based on PhUSE, TransCelerate, HIPAA and CDISC guidelines, generating narratives from the anonymized datasets would be an avenue worth exploring. A minimum requirement with this approach would be to prevent misclassifying clinical data as PPD/PII and removing clinically relevant information during the anonymization process. This approach would aid in the use of defined rules, leading to:

- Consistency of anonymized outputs – datasets and narratives
- Better matching of PPD/PII
- Risk calculation on IPD that could drive the anonymization

However, anonymized data cannot be claimed to be a 100% anonymous, as there is always a possibility of patient re-identification. In this age of big data and cloud computing, it becomes increasingly difficult to eliminate the risk of patient re-identification completely. Therefore, generation of narratives from such anonymized data may add to that risk. However, the risk would be quantifiable at this stage (and not qualitative, as seen in reactive techniques) and could be reduced by streamlining further. A risk assessment of re-identification will need to be conducted. Currently, the tools available for risk assessment are not straightforward. However, with advances in the area of transparency, a number of vendors are collaborating to explore technologies that could gradually address the implementation issues. Proactive anonymization offers a promising alternative to reactive techniques for narrative anonymization. With increased guidance, better defined expectations and availability of more standardized technologies, it would become easier to manage the processes described above and make it more accessible to all sponsors. Part II of Policy 0070 may also open useful avenues to aid in the exploration of better techniques.

Below is a table outlining differences between reactive (mainly redactions) and proactive anonymization of narratives:

Reactive	**Proactive**
Common sense approaches or rules of thumb	Quantitative methods that analyze the data itself to measure the risk of re-identification
Manual manipulations to the text; highly subjective	Structured/consistent manipulation of text
Decreased data utility	Increased data utility

It is important to ensure the same approach is followed for the rest of the CSR body as for the narratives, to avoid the inevitable inconsistencies that would arise if different anonymization approaches were applied.

A Proposed Prospective approach:

Challenges

Primarily the challenge posed by patient narratives is that the amount of data on a patient that can include direct and indirect identifiers that are not easily detected, and are linked to data within the study and external to the study which constitute a patient's data mosaic, or all the data publicly available on a specific person. At some sponsors, narratives are not programmed, but rather are more consistent with medical narratives often used in the clinical setting, containing highly specific and identifiable data elements in a format often difficult to anonymize without use of a NLP tool 12. The unstructured format of the narratives adds another layer of complication. Additionally, EMA enforcement of requirements as part of Policy 0070 has been changing. Initial submissions by sponsors to the agency have allowed large amounts of redaction or complete redaction strategies. However, as sponsors have moved into additional submission packages, the agency has not been as forgiving in allowing a full redaction strategy and is requiring more anonymization techniques to be employed and documentation on risk and data utility.

Considerations

Sponsors that employ a programmatic solution to generating patient narratives are properly positioned to produce prospectively anonymized narratives. However, sponsors do not need to change their approach if they hope to produce anonymized narratives or even reduce risk of identification even when using processing through NLP tools. As such, sponsors changing the generation from a manual process in which a clinician authors the narrative to one that a programmer/tool creates the narrative, is a significant management task change.

Approach

As per International Conference on Harmonization (ICH) E3 (Section 12.3.2), a patient narrative should describe:

- The nature, intensity and outcome of the event
- Clinical course leading to the event
- Timing of study drug administration
- Relevant laboratory measures
- Counter measures
- Action taken with the study drug in relation to the event
- Post mortem findings (if applicable)
- Investigator's and sponsor's opinion on causality

Additionally, patient identifier, age, gender, clinical condition, disease being treated, relevant medical history, concomitant and prior medications should be included.

A general approach to ensure that the above criteria are met, would include the follow steps:

- Mapping sponsor submission data standards to the criteria.
- Develop narrative annotated templates or scripts at the global, therapeutic, compound and indication level where appropriate.
- These narratives with "fill in the blank" sections populated by program code.
- Develop business rules on punctuation, capitalization and sponsor-specific sentence structure.
- Create program code that reads in the submission data, and processes the template into a readable narrative format adhering to sponsor business rules.
- Clinicians may review and provide additional changes to the narratives in a way that can be consumed electronically, and presented in a final version in a clinical discussion section of the narrative.
- Follow sponsor SOPs on validation or qualification.

There are several approaches available and have been presented at SAS user group meetings and other industry conferences. The links providing additional resources on forming a specific approach suitable for sponsors are listed in the Reference section of this document.

Sponsors have to bear in mind that primary recipients of the ICH E3 CSR are reviewers at the national competent authorities. If a sponsor chooses to proactively anonymize a CSR, a careful assessment has to be done to confirm that the document is easily readable and all required information is readily found. If the documents submitted are proactively "over anonymized" they may irritate reviewers and result in clock-stops, subsequent questions and eventual market entry delays. Sponsors should make an informed decision based on their business needs, proactive planning, and availability of tools and resources to meet the requirements of Policy 0070. An alternative approach for prospective CSRs would be to encourage lean writing, so only the required information is presented in the report and the rest is cross-referenced in other documents that are out of scope of Policy 0070 disclosure.

References:

1. Khaled El Emam, 2017 - An Analysis of Anonymization Practices in Initial Data Releases Pursuant to EMA Policy 0070

2. Maund E, Guski LS, Gøtzsche PC, 2017 - Considering Benefits and Harms of Duloxetine for Treatment of Stress Urinary Incontinence: a Meta-analysis of Clinical Study Reports

3. Tammisetti et al, 2017 - A Guide to Programming Patient Narratives

4. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule
5. TransCelerate Biopharma, Inc., 2016 - Protection of Personal Data in Clinical Documents – A Model Approach

6. TransCelerate Biopharma, Inc., 2016 - Data De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach

7. Nikita Raaj, Automated Tool for Anonymization of Patient Records

8. Yang et al, 2015 - Automatic detection of protected health information from clinic narratives

9. Meystre et al, 2010 - Automatic de-identification of textual documents in the electronic health record: a review of recent research

10. The BMJ Opinion - The release of regulatory documents under EMA policy 0070: Now you see them, now you don't

11. Kushida et al, 2012 - Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies

12. Girase et al, 2016 - Programmed Patient Narratives Using SAS®