Paper SA08

Harnessing AI/ML Technologies to Expedite Anonymization of Unstructured Data: Implications for EU CTR

Niamh McGuinness, Privacy Analytics (IQVIA), Ottawa, Canada
Muhammad Oneeb Rehman Mian, Privacy Analytics (IQVIA), Ottawa, Canada

ABSTRACT

The EU Clinical Trials Regulation (CTR) requires the publication of many documents submitted during the clinical trial lifecycle. Personal and commercially confidential information must therefore be protected through redaction and/or anonymization. In an effort to align existing anonymization solutions for clinical study report (CSR) data to the novel demands of EU-CTR, practical challenges emerge such as complex document formatting, high throughput, and varying regulatory requirements. In this regard, Artificial Intelligence and Machine Learning (AI/ML) technologies can be used to successfully build and deploy at scale a system for anonymizing CSR data. Clinical Natural Language Processing (c-NLP) aids in handling unstructured data, in particular clinical documents, to accelerate detection and classification of personal information for anonymization. Here we will describe how EU CTR compares to other disclosure initiatives, and how AIML approaches such as c-NLP can expedite redaction and anonymization to help sponsors navigate the new transparency landscape under EU CTR.

INTRODUCTION TO CLINICAL TRIAL TRANSPARENCY

Clinical trial transparency has been an area of increased focus for clinical trial sponsors over the last two decades, aimed at building public trust in pharmaceutical research, reducing duplication of clinical trials and ultimately improving healthcare outcomes and driving innovation.

Clinical trial transparency can encompass several activities. The first is clinical trial registration in which key information about a clinical trial is shared to a public registry such as *clinicaltrials.gov*. More recently, there has been a greater push towards demystifying language which could be considered technically prohibitive for a general audience to understand, including the prospective participants of clinical trials. As such, plain language or lay summaries make technical language pertaining to clinical trials more accessible to the public and are also considered as a key element of clinical trial transparency. In both registration and plain language summaries, the information presented is intended to summarize findings across entire studies rather than allude to individual participants.

However, there are also some scenarios in which individual participant data must be shared. Clinical trial sponsors can opt to voluntarily share structured, row-level individual participant data with researchers or requestors through dedicated clinical data-sharing platforms such as Vivli, YODA or others. Furthermore, there are also regulatory requirements for the publication of clinical trial documents such as clinical study reports and associated summaries. In both cases, the personal data contained in these structured and unstructured data must be anonymized to protect participant privacy.

REQUIREMENTS FOR CLINICAL DOCUMENT PUBLICATION

This paper aims to present an optimal strategy for handling publication requirements under the EU Clinical Trials Regulation, but first it is important to contextualize this requirement versus other existing requirements for clinical document publication.

The EMA Clinical Data Publication Policy (Policy 0070) came into effect in 2015. It required any clinical trial sponsors seeking a marketing authorization through the centralized procedure in the EU to publish clinical study reports and associated appendices through a dedicated Clinical Data Portal. To prepare the documents for publication, sponsors were required to redact or otherwise anonymize the personal data pertaining to clinical trial participants to reduce their identifiability/re-identification risk to below the threshold of 0.09. While EMA stated a preference for an analytical or quantitative approach to measuring and mitigating this risk, as well as utility preservation via use of generalization or other transformation strategies, a non-analytical or rules-based approach coupled with redaction was also acceptable. Health Canada launched an analogous initiative in 2019 – the Public Release of Clinical Information policy. This built on Policy 0070 in that members of the Canadian public could also request the publication of clinical trial documents pertaining to older trials as well as those eligible for publication from 2019 onwards.

The EU Clinical Trials Regulation entered into force in 2014, however, its application was dependent upon the functionality of a database called CTIS or the Clinical Trials Information System. This went live on January 31st, 2022. While there was a transition period of a year in which clinical trial sponsors could adhere to the old Directive, from 31st January 2023, all new clinical trial applications must be created under the new Regulation. A key differentiating element of this Regulation is many of the documents created and submitted throughout the clinical trial lifecycle must now also be published. Even if the sponsor selects deferral of publication, both the original and redacted or anonymized versions

of the documents must be uploaded at the same time. Given the turnaround time for these uploads can be as few as 12 calendar days, sponsors must be prepared and have a robust, scalable, and repeatable strategy in place to protect personal data in these documents.

INTRODUCTION TO ANONYMIZATION

Specific guidance varies between regulators, but broadly speaking, we can consider anonymization to mean rendering data subjects non-identifiable/indistinguishable through the transformation or removal of identifying elements of personal information**.** There are three elements to creating an appropriate anonymization strategy:

1. What to include: What type of information should be included in the assessment of identifiability or re-identification risk?
2. How to assess? Which method of identifiability assessment should be used? A quantitative/analytical approach or a qualitative/non-analytical approach?
3. How to mitigate? Once we have assessed identifiability, what can we do to reduce it sufficiently?

What to include: There are usually two categories of data subjects that we encounter in clinical documents – clinical trial participants and sponsor or site personnel. Identifiable information pertaining to trial participants can be classified into direct and indirect identifiers. Direct identifiers are information that can uniquely identify an individual; some examples of direct identifiers include Subject IDs, email addresses, and names. Indirect identifiers are elements of personal information that cannot uniquely identify an individual, but which may be able to do so in combination. Indirect identifiers are typically the greatest contributor toward identifiability and are also more analytically useful. By comparison, information pertaining to sponsor and site personnel tends to fall exclusively under the direct identifier header.

How to assess and mitigate: There are two main methods through which identifiability can be assessed, qualitative and quantitative. A qualitative approach is rules-based and non-analytical. This type of approach is usually driven by observations and assessments by responsible parties. Especially when coupled with redaction, this type of approach can lead to near-complete removal of personal information, as the absence of a statistical measure can make retention of information difficult to justify. This can have a negative impact on data utility.

On the other hand, a quantitative approach to identifiability assessment is statistically derived and analytical. It is usually driven by a mathematical model of identifiability. A key advantage of a quantitative approach is that transformations (which may encompass techniques like generalization and date offsetting) can be adjusted based on the numeric measure of identifiability. In this way statistical anonymization can preserve utility where possible whilst providing assurance that identifiability is sub-threshold.

AI/ML AND ANONYMIZATION

AI/ML-driven statistical anonymization is a powerful approach to detect, quantify, and transform the identifiable elements within clinical trial documents. Machine Learning techniques like clinical natural language processing (c-NLP) can detect and extract personal information within unstructured documents based on the context in which they arise, at speeds much faster than a human analyst is able to perform.

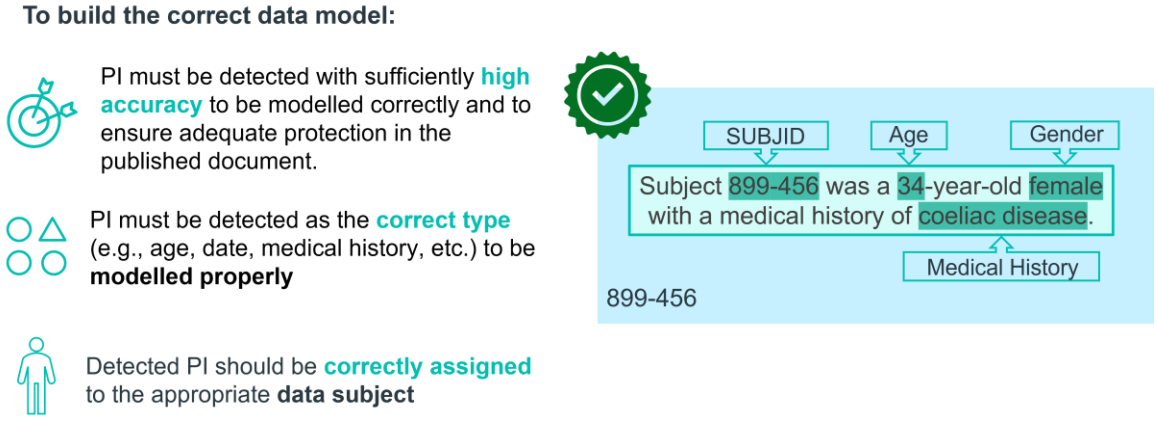

Figure 1 :The challenges of c-NLP-driven statistical anonymization of clinical trial documents

However, a c-NLP-enabled approach to statistical anonymization of clinical trial documents does present several challenges, as outlined in Figure 1. To build the correct data model, personal information must be detected with sufficiently high accuracy to be modelled correctly and to achieve adequate protection in the published document. Personal information must also be detected as the correct variable type (e.g., age versus medical history versus date,

etc.) to be modelled properly. Finally, detected personal information should be associated with the correct data subject to ensure its contribution to an individuals' identifiability is being accurately assessed.

Given these challenges, the use of qualitative redaction to transform documents may substantially simplify the path to compliance for EU Clinical Trials Regulation, as personal information need only be captured in absolute terms – there is no requirement for the information to be modelled correctly or to be associated with the relevant participant. Instead, the primary requirement is that the personal information detection process is accurate enough to justify a determination that data subjects are rendered anonymous.

REDACTION VERSUS STATISTICAL ANONYMIZATION FOR EU CTR

The determination of whether statistical anonymization or redaction is most appropriate for transparency is ultimately within the purview of the trial sponsor. Especially for Part 1 and Part 2 documents that are part of the clinical trial application, which must be uploaded quickly, vary in size and format, and in which company confidential information (CCI) will constitute a greater concern for sponsors, redaction is likely to be the best choice. Clinical reports, which may also be eligible for publication under other transparency initiatives could be subject to statistical anonymization but will also likely be redacted. We anticipate that AI/ML-driven approaches can be a more efficient and viable option for transparency purposes.

TESTING OUR REDACTION APPROACH ON EU CTR DOCUMENTS

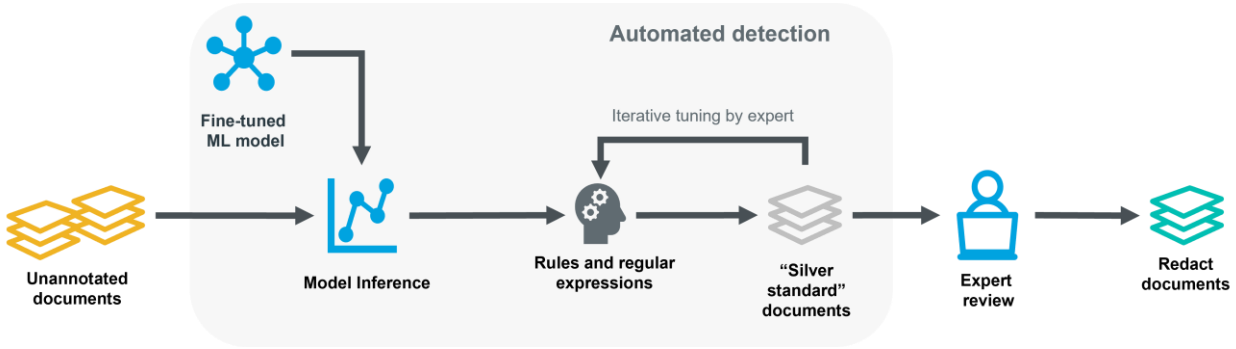

Figure 2 :c-NLP-driven PI redaction workflow for EU CTR

We have found that a hybrid model combining a machine learning-based detection with a rules-based approach tends to be most effective for detection personal information in clinical text. With the former we could leverage large, pre-trained language models for contextual understanding of text, and fine-tune on real-world data from completed anonymization projects. With the latter we can optimize the c-NLP-based detection through application of regular expressions to detect attributes with known formatting, and then iteratively improve rules to target efficiency gains with each submission. This process concludes with an expert review, leading to documents which have been redacted at speed and scale with a high degree of accuracy.

RESULTS

In this table we have outlined the documents tested, the average length of such documents and the personal data detected therein.

Document	Description	Page Counts	Personal Data Detected
Investigator Brochure	Summarizes the body of information about an investigational product obtained during a trial	100-300 pages	Participant/case numbers, ages, countries, medical histories
Protocol	Outlines the rationale, objective(s), methodology used, and analysis plan for a clinical study	100-300 pages	Site and sponsor personnel names, contract details and addresses.
Investigator CVs	Presents investigator eligibility and experience	2-5 pages	Principal investigator names, contact details and addresses

Clinical Study Reports	Provides detail (including individual participant data) about the safety and efficacy results of a clinical trial.	100-100,000 pages (with appendices)	Rich participant data like subject IDs, age, important calendar dates, demography and vital signs, medical histories and adverse event information.

Table 1: Key information about the clinical trial documents tested

In addition to the detail outlined above, we observed the following:

- Sections of text with sparse personal information tend to be best handled through manual redaction.
- C-NLP model predictions tend to be most effective when personal information has contextual dependence such as in clinical narratives (typically contained in clinical study reports but also in IBs.
- Writing custom regular expressions tended to be most efficient for capturing personal information in tabular data.
- Time spent manually correcting documents tended to vary more strongly with document length than density of personal information.

From a quantitative perspective, we generated the following metrics:

- In a particular CSR narrative format, the AI/ML model achieved 100% personal information detection accuracy (evaluated by a human annotator)
- Use of custom regular expressions for documents with predictably formatted personnel-related information reduced manual effort by up to 50%.
- Using an c-NLP-driven PI detection process showed a 20% improvement in accuracy relative to a fully manual approach on documents like the Investigator Brochures.

CONCLUSION

In conclusion, EU Clinical Trials Regulation has substantially increased the scope of public clinical document releases required of clinical trial sponsors. When evaluating the types of documents which will be eligible for disclosure under this Regulation, we found that a c-NLP-driven approach allowed for successful redaction of documents in scope. Given the increased throughput requirements, c-NLP-driven redaction is a viable option to satisfy regulatory criteria.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Niamh McGuinness
Privacy Analytics
251 Laurier Avenue West, Suite 1000
Ottawa, K1C 1G1
+1-613-297-8330
nmcguinness@privacy-analytics.com