Paper ML21

Strategic Authoring of Clinical Study Reports to Leverage AI for Data Privacy and Public Disclosure

Anaya Rehman, Certara, Canada

ABSTRACT

Since the push towards transparency of clinical trials with the establishment of clinical trial registries, sponsors face increasing challenges in ensuring timely regulatory compliance. This increasing demand for public disclosure necessitates a strategic approach to data protection. Personal and company information is often embedded in lengthy clinical documents which makes its accurate capture a laborious process. AI-driven anonymization offers a scalable solution for safeguarding information. However, the AI tools' effectiveness depends upon structured and standardized document authoring. By implementing proactive standardized authoring through Strategic Medical Authoring for Regulatory Transparency (SMART) approach, sponsors can optimize clinical document development to improve automation in anonymization. This approach uses the regulatory understanding of the guidance to devise best practices in medical writing with future disclosure in mind. This paper explores the interdependence of structured authoring, automation, and AI-driven data protection, demonstrating how a proactive approach to clinical document development enhances compliance, reduces manual effort, and facilitates efficient public disclosure while upholding the highest standards of data privacy.

Keywords: Strategic authoring, AI-driven anonymization, public disclosure, data privacy.

INTRODUCTION

Clinical trial transparency has evolved significantly over the past two decades, with registries and public disclosure requirements becoming a standard part of the regulatory landscape. Regulatory agencies have implemented policies such as Health Canada's Public Release of Clinical Information and European Medicines Agency's (EMA's) Policy 0070, mandating the disclosure of clinical trial data, including clinical study reports (CSRs) and other regulatory documents, to enhance public trust and scientific reproducibility[1,2]. However, this push for openness comes with the responsibility of protecting personal and commercial confidential information. Given the complexity and length of regulatory submissions, identifying and redacting personal and commercial information is a time-intensive and resource-heavy task[3].

A crucial factor for success is designing and authoring clinical documents with public disclosure as an ultimate step of the process. Achieving this requires proactive collaboration between medical writers and transparency and disclosure experts. This paper seeks to bridge the gap between these roles by introducing an innovative approach—Strategic Medical Authoring for Regulatory Transparency (SMART)—which optimizes the authoring process to improve efficiency in the anonymization of regulatory clinical trial documents.

SMART is a strategic solution that integrates regulatory expertise with medical writing best practices to enhance document clarity while anticipating future disclosure needs. By aligning clinical content with regulatory expectations from the outset, the SMART approach facilitates efficient data protection and enables the application of artificial intelligence (AI) tools for enhanced automation in safeguarding sensitive information. AI-driven techniques, such as natural language processing (NLP), have demonstrated promise in automating the identification of confidential data, thereby streamlining compliance efforts[4].

PROTECTION OF PERSONAL DATA IN CLINICAL STUDY REPORTS

A CSR is a comprehensive document that details various aspects of a clinical trial, including its methodology, statistical analyses, and results. These reports often contain personal data, requiring compliance with stringent data

privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA)[5] in the United States and the General Data Protection Regulation (GDPR) in the European Union[6]. HIPAA Privacy Rule protects all *'individually identifiable health information'* that can be used to reasonably identify an individual. It defines 18 specific personal identifiers classified as protected health information (PHI). These include names, geographic details below the state level, specific dates associated with an individual (excluding the year), contact details, social security numbers, medical record numbers, biometric data such as fingerprints, and other unique identifying characteristics[5]. Similarly, under GDPR, personal data is defined as any information linked to an individual, '*data subject*', that enables direct or indirect identification. This includes identifiers such as names, identification numbers, location data, online identifiers, or attributes related to physical, physiological, genetic, mental, economic, cultural, or social identity[6].

Non-compliance with these regulations can lead to significant legal penalties, underscoring the need for robust anonymization strategies before public disclosure. Techniques such as pseudonymization, generalization, suppression, and randomization play a crucial role in anonymizing sensitive information, ensuring compliance with regulatory requirements while maintaining data utility for secondary use[7].

The SMART approach is closely aligned with the need to anonymize these identifiers. By focusing on the inclusion of only necessary content and structuring documents in a way that isolates personal data, the SMART approach standardizes documents and streamlines the anonymization process. The result is a fast, efficient, and more consistent compliance with regulatory standards without compromising the data utility of clinical documents. The overarching goal is to facilitate the public disclosure of clinical data while ensuring the protection of personal privacy.

ROLE OF AI IN CLINICAL DATA PROTECTION

Modern AI-based anonymization software offer a solution by automating the identification and removal of sensitive information from clinical documents, ensuring compliance with data protection regulations such as GDPR or HIPAA. AI models are trained, validated, and deployed in real-world clinical environments to enhance data privacy without compromising data utility (See Figure 1).

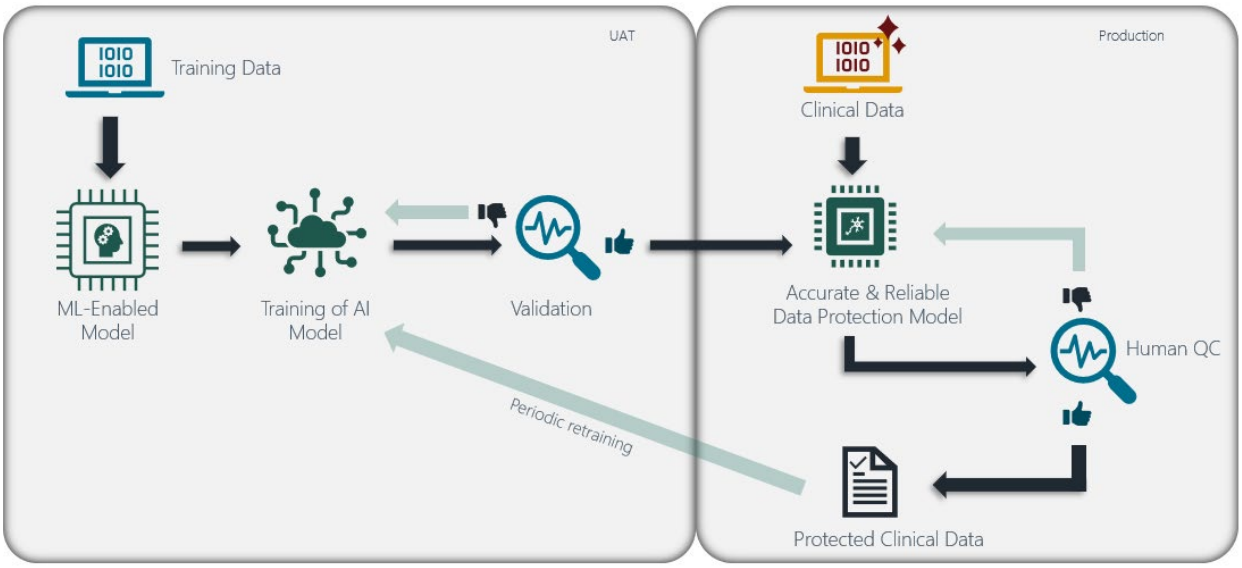

Figure 1. AI Model Workflow for Clinical Document Anonymization.

In the context of data protection, User Acceptance Testing (UAT) with synthetic data that simulates real-world data, is used to ensure the software's effectiveness and privacy compliance before deployment with real patient information. The process begins with training an AI model on large datasets of clinical data, where personal identifiers like subject IDs, demographic information, and addresses are labeled. Through iterations, the AI model learns to recognize patterns associated with sensitive information and refines its ability to anonymize data accurately. To ensure that the model functions as expected to identify and anonymize data, validation is achieved through iterative refinement, performance metrics (such as precision ad recall) and human-in-the-loop validation.

By using fake data during UAT, organizations can mitigate the risk of data breaches, re-identification, or non-compliance, as the process ensures the software can handle real-world data securely. It provides an opportunity for stakeholders to verify that the anonymization software meets both regulatory standards and operational needs before it is used on actual clinical data.

Once the AI model is trained and validated as a reliable data protection model, it is deployed into real-world environments where it processes sensitive medical documents automatically. To ensure continuous accuracy, AI-based anonymization tools operate with a feedback loop, where human reviewers assess anonymized documents for errors and refine them as needed. Furthermore, to maintain high accuracy and compliance, the AI model undergoes re-training at a regular cadence. During this periodic retraining, new labeled datasets, including updated medical terminology and emerging patterns in clinical documentation, are used to refine the model. This continuous improvement ensures the AI can adapt to evolving data formats, maintain effectiveness in anonymizing sensitive information, and meet changing regulatory requirements.

Standardization in authoring clinical documents plays a crucial role in enhancing the effectiveness of AI-based anonymization tools. When clinical data is consistently structured and formatted according to standardized guidelines, AI algorithms can more easily and accurately identify and remove personal identifiers, improving both the speed and reliability of the anonymization process. Additionally, standardized formats reduce the ambiguity in how information is presented. This clarity allows to leverage the available AI tools better to recognize personal information.

STRATEGIC MEDICAL AUTHORING FOR REGULATORY TRANSPARENCY (SMART)

To fully grasp the necessity of the SMART approach, it is essential to first understand the ultimate result - successful public disclosure (See Figure 2). Achieving this outcome requires stringent data protection measures, which, in turn, demand effective automation. However, for automation to function optimally, standardized authoring practices must be in place. This interdependence highlights the fundamental role of SMART in ensuring a seamless, compliant, and efficient disclosure process.

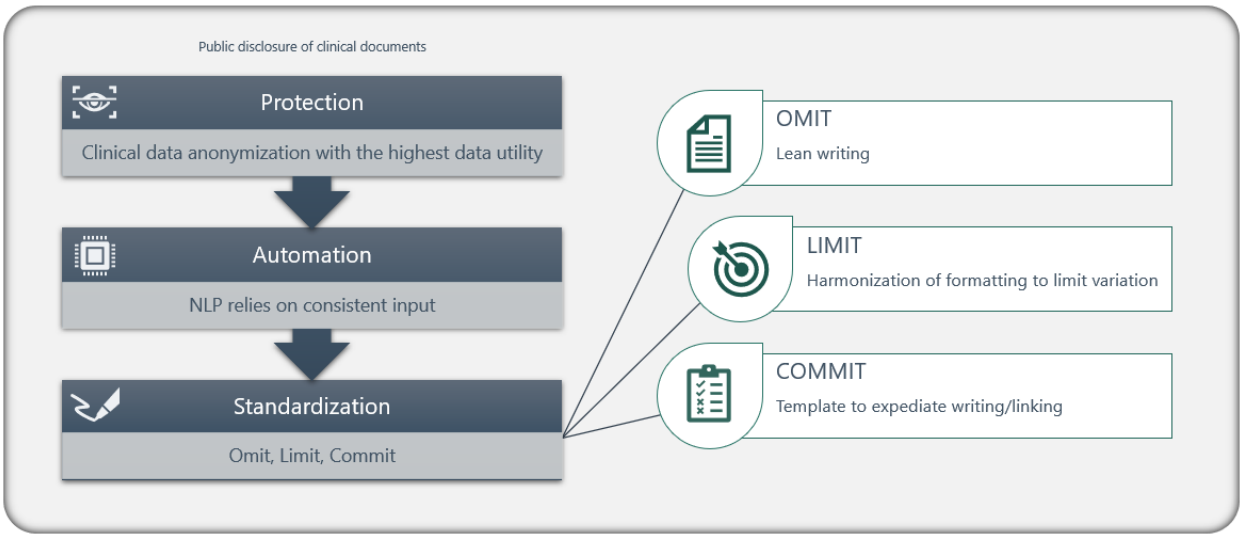

Figure 2. Interdependence of Standardization, Automation, and Data Protection in Achieving Public Disclosure

It is widely recognized that adherence to established standards in clinical document authoring such as the ICH E6 guidelines, enhances the quality of regulatory documents. According to the American Medical Writers Association (AMWA), the most common issues in clinical documents are excessive length, verbosity, and repetition[8]. A structured, templated approach enhances the efficiency of document anonymization. It not only simplifies regulatory compliance but also expedites the approval process, ultimately reducing the time needed to obtain marketing authorization. The SMART approach emphasizes this streamlined standardization in document creation through several key strategies:

- **Omit** non-essential details from clinical reports, such as excessive confidential business information or unnecessary social context in narratives. This ensures that documents remain focused on essential scientific content, making them more concise and easier to interpret.
- **Limit** variations in personal data presentation by applying uniform formatting, which enhances clarity and coherence across documents.
- **Commit** to standardized terminology and predefined structural templates to optimize AI-driven identification and protection of sensitive information. This structured approach enables efficient compartmentalization, making it easier to locate and anonymize personal data.

By strategically authoring documents with these principles, unnecessary information is minimized, ensuring they are disclosure-ready while adhering to regulatory requirements and protecting individual privacy.

To further explore this concept, it is important to understand its connection to personal identifiers and how medical writers can apply best practices. Personal identifiers are generally categorized into two types: direct and indirect. Direct identifiers uniquely distinguish a clinical trial participant, while indirect or quasi-identifiers, as the name implies, can lead to identification when combined with other information. Let's consider an example of a direct identifier, such as a subject ID. For example, in a clinical trial, participant identifiers might be formatted inconsistently across various documents—sometimes written as participant ID, sometimes as 'Patient ID,' or even with different numbering systems. If one document refers to a participant as 'P123' and another refers to the same person as 'Patient_123,' it could create confusion when linking information from different sources.

The SMART approach addresses this by promoting uniformity in how identifiers are handled. For instance, by consistently using a Universal Subject ID (USUBJID) across all documents, such as 'CT123-001-1001' for a participant in study CT123 at site 001, there is no room for confusion. When authoring clinical documents, it's important to **omit** unnecessary details related to the subject ID, such as redundant information that don't contribute to the scientific or regulatory context. For example, additional personal details like an address should be omitted to prevent over-exposing personal information. The next step is to **commit** to a standardized format for the USUBJID across all clinical documents. By consistently using the same format, such as 'CT123-001-1001,' the anonymization process becomes more efficient and less prone to error. This uniformity ensures that sensitive data can be identified and protected more easily through automation. And finally, it's crucial to **limit** variations in how the USUBJID is presented across documents. If one document refers to a subject ID as '1001' and another as 'CT123-001-1001,' this inconsistency could create confusion. The issue becomes even more complex when multiple studies are involved, such as extension studies with overlapping participants, or when similar subject IDs exist for more than one person, for instance, '1001' for CT123 and '1001' for CT456, leading to potential misidentification and data integrity challenges. By limiting these variations, the subject ID is always presented the same way, which helps in streamlining data processing, linking information correctly, and enhancing the overall anonymization process. This consistency makes it easier to automate anonymization, reduces errors, and ensures that personal data is properly protected while maintaining the integrity of the clinical data.

CONCLUSION

As AI continues to evolve, data anonymization tools will improve. AI's role in document authoring itself will also expand, making it increasingly possible to automate the writing process as well. This would not only improve the efficiency of document creation but also contribute to maintaining high standards of accuracy, compliance, and privacy protection. Furthermore, emerging trends in AI model refinement are expected to revolutionize how clinical documents are authored, anonymized, and verified.

Looking to the future, there are several key recommendations to improve the clinical document authoring process. First, organizations should review current CSR templates to identify areas where standardization can be added. This could involve creating more uniform structures for participant demographics, medical histories, and adverse event reporting, making it easier to apply automated anonymization techniques consistently across documents. Next, incorporating stakeholder collaboration throughout the document authoring process is critical. This means involving data scientists, regulatory affairs teams, and clinical researchers early to ensure templates and anonymization practices meet both technical and regulatory requirements. Additionally, creating a feedback loop for continuous improvement based on stakeholder input will allow the process to evolve as new standards emerge. Lastly, investing in regular training for teams to stay updated on the latest advancements in AI, regulatory standards, and anonymization best practices is essential. Encouraging interdisciplinary collaboration and knowledge sharing can help enhance the overall efficiency and compliance of clinical documentation, ultimately leading to more accurate, privacy-secure, and timely submissions.

In conclusion, the use of AI-driven anonymization tools in clinical data processing offers a powerful solution for protecting sensitive patient information while maintaining the integrity and usability of the data. The SMART approach further enhances this process by promoting streamlined standardization in document authoring, ensuring that clinical data is disclosure-ready, consistently anonymized, and compliant with regulatory requirements, all the while safeguarding personal privacy.

REFERENCES

1 European Medicines Agency (EMA). (2019). EMA Policy 0070 on publication of clinical data for medicinal products for human use. Retrieved from https://www.ema.europa.eu

2 Health Canada. (2019). Public release of clinical information. Retrieved from https://www.canada.ca

3 Wieseler B, McGauran N, Kerekes MF, et al. (2019). Impact of clinical trial transparency on data integrity and trust. BMJ, 367, l5891.

4 Bielinski SJ, Bushell T, Mosley JD, et al. (2021). Natural language processing for the identification of sensitive data in clinical documents. Journal of Biomedical Informatics, 118, 103799

5 HHS (U.S. Department of Health and Human Services). (2020). Summary of the HIPAA Privacy Rule. Retrieved from https://www.hhs.gov

6 European Commission. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council—General Data Protection Regulation (GDPR). Retrieved from https://eur-lex.europa.eu

7 El Emam K, Mosquera L, Bass J. Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. J Med Internet Res. 2020 Nov 16;22(11):e23139. doi: 10.2196/23139. PMID: 33196453; PMCID: PMC7704280.

8 Cooper J, Chamberlain James L, Affleck J, Bass B, Forjanic Klapproth J, Harris D. Value of Medical Writing: The Regulator's Perspective. AMWA. 2021;36(4). doi:10.55752/amwa.2021.83

CONTACT INFORMATION

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

Author Name: Anaya Rehman
Company: Certara
Address: 4 Radnor Corporate Center suite 350, Radnor, PA 19087, USA
Work Phone: +1 514-395-7729
Email: anaya.rehman@certara.com
https://www.linkedin.com/in/anaya-rehman-182bbb86/