Paper PP25

Optimizing AI Software for Automated Protection of Personal and Commercially Sensitive Data

Honz Slipka, Certara

Abstract

Ensuring clinical data transparency while protecting sensitive patient information has become a major challenge due to the vast volume and variability of data in pharmaceutical research. Traditional anonymization methods often lack consistency and efficiency, increasing the risk of data breaches and regulatory non-compliance. To address these issues, integrating artificial intelligence (AI) and machine learning (ML) offers a transformative solution by automating anonymization processes, improving accuracy, and reducing human error. The SMART approach developed by Certara, along with the adoption of the hybrid methodology, promotes proactive data protection by embedding anonymization strategies early in drug development and leveraging AI automation. By doing this, the pharmaceutical industry can achieve secure data sharing while safeguarding patient privacy, ensuring compliance, and mitigating financial and reputational risks.

Introduction

In an era where clinical data transparency is crucial for regulatory submissions and public disclosure, ensuring robust anonymization techniques has become a pressing challenge. The pharmaceutical industry is dealing with an unprecedented volume and variability of sensitive data, with millions of data points per dossier requiring meticulous protection against de-anonymization and re-identification. Traditional anonymization methods often struggle with consistency, efficiency, and accuracy, leading to prolonged processing times and potential data breaches.

To address these challenges, a proactive and technology-integrated approach to data protection is essential. Leveraging artificial intelligence (AI) and machine learning (ML) offers a transformative solution by automating anonymization processes, improving accuracy, and reducing human error. AI-enabled models can intelligently predict which data points require protection, streamline clinical privacy strategies, and establish a standardized framework for anonymization. The integration of AI not only enhances the speed of processing protected personal data (PPD) and commercially confidential information (CCI), but also ensures compliance with evolving regulatory standards, such as the European Medicines Agency (EMA) Policy 0070 and Health Canada's PRCI.

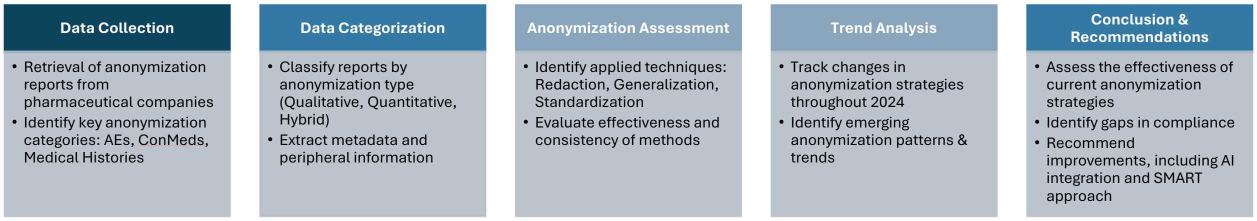
The Certara-developed SMART approach (Strategic Medical Authoring for Regulatory Transparency) emphasizes upstream data protection by embedding anonymization strategies early in the drug development lifecycle. This proactive methodology ensures that sensitive data considerations are addressed from the outset, minimizing risks while maintaining transparency. With AI-driven automation, the pharmaceutical industry can achieve disclosure without exposure—balancing the need for open clinical data sharing with the imperative of protecting patient privacy. As regulatory requirements evolve and data breaches pose increasing financial and reputational risks, AI-driven clinical data anonymization represents the future of secure and efficient data protection.

Methodology

Data was collected from the EMA Policy 0070 portal to analyze anonymization strategies applied to clinical documents, assess emerging trends and review acceptable regulatory standards in order to develop the most practical approach to anonymizing clinical data. The dataset comprises regulatory reports from various pharmaceutical companies, detailing how patient information is protected during public disclosure. The collection process focused on identifying anonymization patterns for key clinical identifiers, including adverse events, concomitant medications and medical histories.

Each regulatory submission was reviewed for qualitative and quantitative anonymization techniques. The primary sources of data were publicly available anonymization reports linked through the EMA portal. Reports were categorized based on anonymization methods such as redaction, generalization, and standardization. The dataset includes metadata such as procedure numbers, marketing authorization holders, and regulatory authorities involved (EMA or Health Canada).

This structured collection methodology enabled a comparative analysis of anonymization strategies across different regulatory submissions, providing insights into industry trends and best practices in clinical data protection.



Data Collection

- Retrieval of anonymization reports from pharmaceutical companies
- Identify key anonymization categories: AEs, ConMeds, Medical Histories

Data Categorization

- Classify reports by anonymization type (Qualitative, Quantitative, Hybrid)
- Extract metadata and peripheral information

Anonymization Assessment

- Identify applied techniques: Redaction, Generalization, Standardization
- Evaluate effectiveness and consistency of methods

Trend Analysis

- Track changes in anonymization strategies throughout 2024
- Identify emerging anonymization patterns & trends

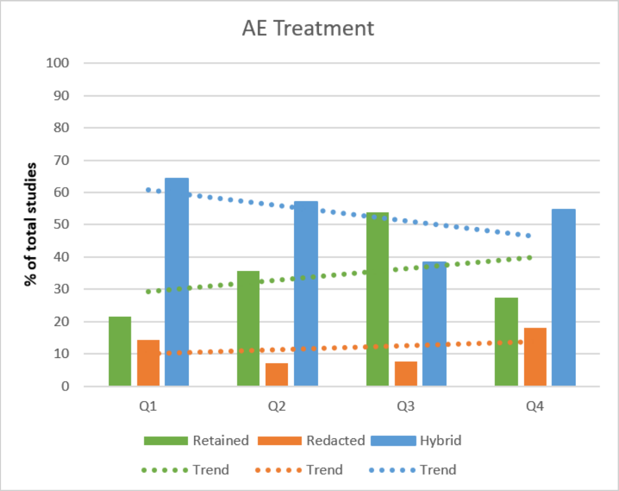
Conclusion & Recommendations

- Assess the effectiveness of current anonymization strategies
- Identify gaps in compliance
- Recommend improvements, including AI integration and SMART approach

Results

The systematic review of subjectively identifying data (adverse events, concomitant medication, medical histories) published in the EMA policy 0070 portal showed numerous findings, patterns and results from the first year of submissions.

Adverse Events



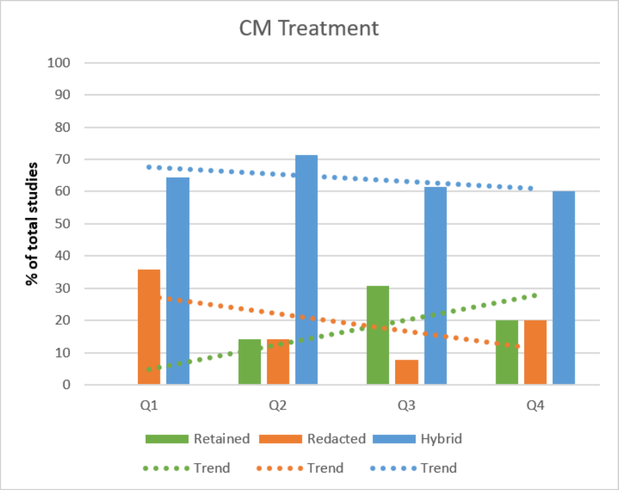
Of the 52 initial marketing authorization studies containing a dverse events posted to the portal, only 34.6% (18) of the studies had retained adverse events, 11.6% (6) of studies had fully redacted adverse events, and 53.8% (28) of studies had some sort of hybrid approach to protecting adverse events information within the submission.

From the hybrid approach studies, notable reasons for subject anonymization and selective redaction were cases in where sensitive information was provided for patients and participants. Sensitive information was often defined as identifiers which were visibly identifiable, clinically rare and often affecting small numbers of the general population, newsworthy information such as traffic accidents or violent criminal activity, or information which could have negative association and could harm or damage participant reputation such as mental health events, sexually transmitted diseases and substance abuse events.

Anonymization of adverse events was mainly through the hybrid approach across all 4 quarters of the year, however, an increase in the number of adverse events being retained went up significantly over the course of the year.

Concomitant Medications

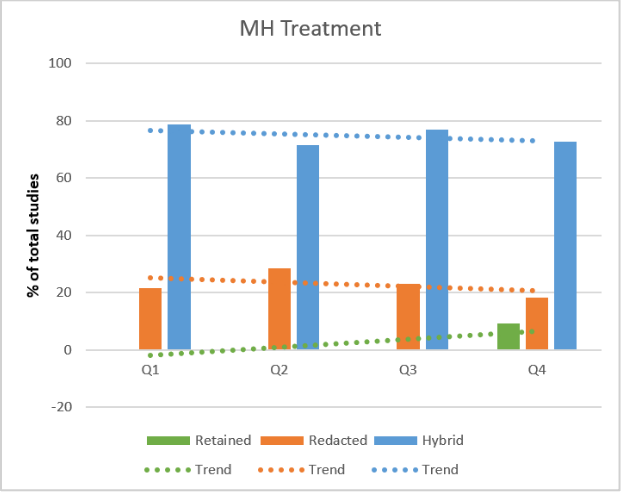
When reviewing the 51 initial marketing authorization studies which contained information related to concomitant medications, 15.7% (8) submissions contained concomitant medications that were fully retained, 19.6% (10) included fully redacted concomitant medications and 64.7% (33) submissions used a hybrid approach to redact selected information in order to protect the identities of the participants involved.



Once again, the hybrid approach was the most common methodology employed to protect the concomitant medications used by the participants within the study. Selective redaction and anonymization of these data was attributed to protecting drugs names and medications which could re-identify individuals within the submissions. Medications included in the list of redacted information were ones that were sex-specific and could therefore identify the sex of individuals, rare and uncommon medications within the general population, and medications that were sensitive and/or related to diseases which could have a negative association with the patients taking them such as medications related to treatment of mental health disorders, sexually transmitted diseases and substance abuse medications.

The most common and most stable strategy of anonymizing concomitant medications throughout the year was the hybrid approach, with an increase of retained concomitant medications near the end of the year, and an equal and average decrease in redacted concomitant medications towards the end of the year.

Medical Histories

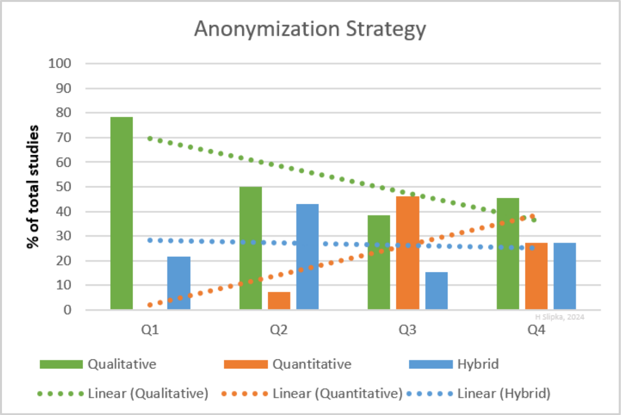


Lastly, when reviewing the treatment of medical histories within the studies posted to the Policy 0070 portal, 1.9% (8) of the 52 studies posted had medical histories completely retained, 23.1% (12) had medical histories completely redacted and 75% (39) of studies had some kind of hybrid approach to protecting medical histories.

Similarly to the other two identifiers, medical histories which were not directly redacted or retained, were treated based on the perceived sensitivity of the history. Medical history terms that were selectively redacted often related to information that could be harmful to the participants, such as data that could harm their employability, their reputation, their insurability, self-esteem or information that could result in a loss of income. Additionally, visibly identifiable medical histories, as well as medical histories that were rare, newsworthy or had a negative association were also often protected.

Treatment of medical histories was consistent throughout the year, with a slight increase in the number of retained medical histories in the 4th quarter of the year, and a minor decrease in the number of studies directly redacting medical histories throughout the year.

Anonymization Methodology



Overall, **the most common anonymization strategy across the board is still the qualitative approach**. This method employs mainly redaction-based data anonymization with little to no empirical risk calculation. Of the 52 studies posted to the portal, 53.8% (28) used the qualitative approach, 19.2% (10) used the quantitative approach, and 27% (14) used a hybrid approach.

While the number of submissions were relatively equal throughout the four quarters (14, 14, 13 and 11 respectively), the number of submissions using a quantitative approach rose as the year went by. Studies using the quantitative approach decreased through the year, and the hybrid approach was the most consistent approach throughout the year.

Category	Total Studies Reviewed	Fully Retained (%)	Fully Redacted (%)	Hybrid Approach (%)	Key Observations
Adverse Events	52	34.6% (18)	11.6% (6)	53.8% (28)	Sensitive AEs were redacted if rare, newsworthy, or reputation-sensitive; increase in retained AEs over time.
Concomitant Medications	51	15.7% (8)	19.6% (10)	64.7% (33)	Hybrid approach was dominant; selective redaction applied to rare, gender-specific, or

					sensitive medications.
Medical Histories	52	1.9% (1)	23.1% (12)	75% (39)	Hybrid approach most common; redactions targeted rare conditions, employability risks, and reputation-sensitive histories.
Anonymization Strategies	52	-	-	-	53.8% Qualitative, 19.2% Quantitative, 27% Hybrid approach; increasing shift towards hybrid strategies.

Discussion

Emerging Trends & Patterns

Throughout 2024, regulatory submissions to the EMA Policy 0070 portal demonstrated significant shifts in anonymization strategies for clinical data. A clear majority of sponsors are still favouring **qualitative anonymization**, with the majority of studies relying on redaction and generalization techniques to protect sensitive patient information. This method was prevalent in handling adverse events, medical histories, and concomitant medications, where direct redactions were frequently applied to eliminate re-identifiable details.

However, a notable increase in **quantitative anonymization** methods was observed, especially in submissions that required risk-based assessments. Techniques such as k-anonymity and data perturbation were employed to balance data protection with usability. Despite its advantages, quantitative anonymization alone was insufficient for handling nuanced data elements like complex medical histories and rare adverse events. Additionally, the cost, resources required, and the time it requires to perform statistical risk calculations makes this process unappealing to many sponsors.

The **hybrid approach**—a combination of qualitative and quantitative methods—emerged as the dominant trend in anonymization strategies. Regulatory reports highlighted a move toward blending direct redactions with statistical assessments to ensure robust data protection while preserving the integrity of clinical datasets for research purposes.

The Hybrid Approach: The Future of Clinical Data Anonymization

As regulatory bodies and the pharmaceutical industry strive for a balance between transparency and patient confidentiality, the **hybrid anonymization approach** has proven to be the most effective. This method integrates the ease and speed of qualitative techniques (such as redaction and generalization) with the empirical and robust statistical support of quantitative approaches (such as risk assessments and algorithmic data modification). The hybrid model allows for a nuanced approach to data protection, ensuring that:

- Highly sensitive or uniquely identifying information is redacted or generalized.
- Broader datasets remain usable for research and analysis.
- Risk-based evaluations guide decisions on subjective data retention versus removal.

Regulatory feedback in 2024 focused heavily on retaining certain identifiers such as adverse events and sex of participants, creating a need for statistical risk-assessment of releasing this information. Blending partial statistical anonymization with rules-based redaction of remaining identifiers creates a submission aligned with regulatory expectations and robust personal data protection. Given the increasing complexity and volume of clinical data, the hybrid approach is becoming the industry standard for anonymization, offering efficiency, flexibility and regulatory compliance.

Leveraging AI and the SMART Approach for Advanced Anonymization

The rapid expansion of clinical data necessitates automation in anonymization processes, making **AI and ML critical tools** for data protection. AI-driven models can predict which data points require protection, automate redactions, and ensure consistency across large datasets. These technologies significantly reduce processing time, enhance accuracy, and minimize human error in anonymization processes.

The **SMART approach** is integral to this evolution. By embedding anonymization considerations early in the drug development process, the SMART approach ensures:

- **Systematic** application of anonymization rules across all clinical study reports and associated datasets.
- **Measurable** assessments to quantify re-identification risk of identifiers in order to minimize redactions and maximize data utility.
- **Automation** to streamline data protection, reducing reliance on manual intervention, especially at early stages of identifying sensitive data in-text.
- **Risk-based** decision-making to determine the level of anonymization required to maximize data utility and transparency.
- **Transparency** in regulatory submissions, fostering trust in data-sharing initiatives.

By integrating AI and the SMART approach, the industry can achieve efficient, scalable, and reliable anonymization, ensuring compliance while maintaining the scientific value of clinical data. The future of data protection lies in automation-driven, hybrid anonymization models that adapt to evolving regulatory and research demands.

Conclusion

AI integration in clinical trial disclosure significantly enhances accuracy, consistency, and processing efficiency. By leveraging **standardization, automation, and advanced data protection**, AI-driven systems create structured frameworks that facilitate regulatory compliance and data integrity. Standardization ensures that anonymization techniques are applied uniformly across datasets, reducing variability and human error. Automation accelerates the anonymization process, enabling timely and efficient data release while maintaining patient confidentiality. Enhanced data protection mechanisms, guided by AI algorithms, mitigate re-identification risks and safeguard sensitive information.

These improvements foster **trust** within the pharmaceutical industry by ensuring that clinical data is both accessible and securely anonymized. AI-powered anonymization not only streamlines regulatory compliance but also supports meaningful data sharing, enabling researchers to derive valuable insights while upholding ethical data handling standards. As the pharmaceutical landscape continues to evolve, AI-driven approaches will be crucial in maintaining a balance between transparency and privacy, reinforcing the industry's commitment to responsible data disclosure.

Contact Information

Honz Slipka
Certara
+1 514-395-1763
Honz.slipka@certara.com
https://www.certara.com/regulatory-science/transparency-disclosure/