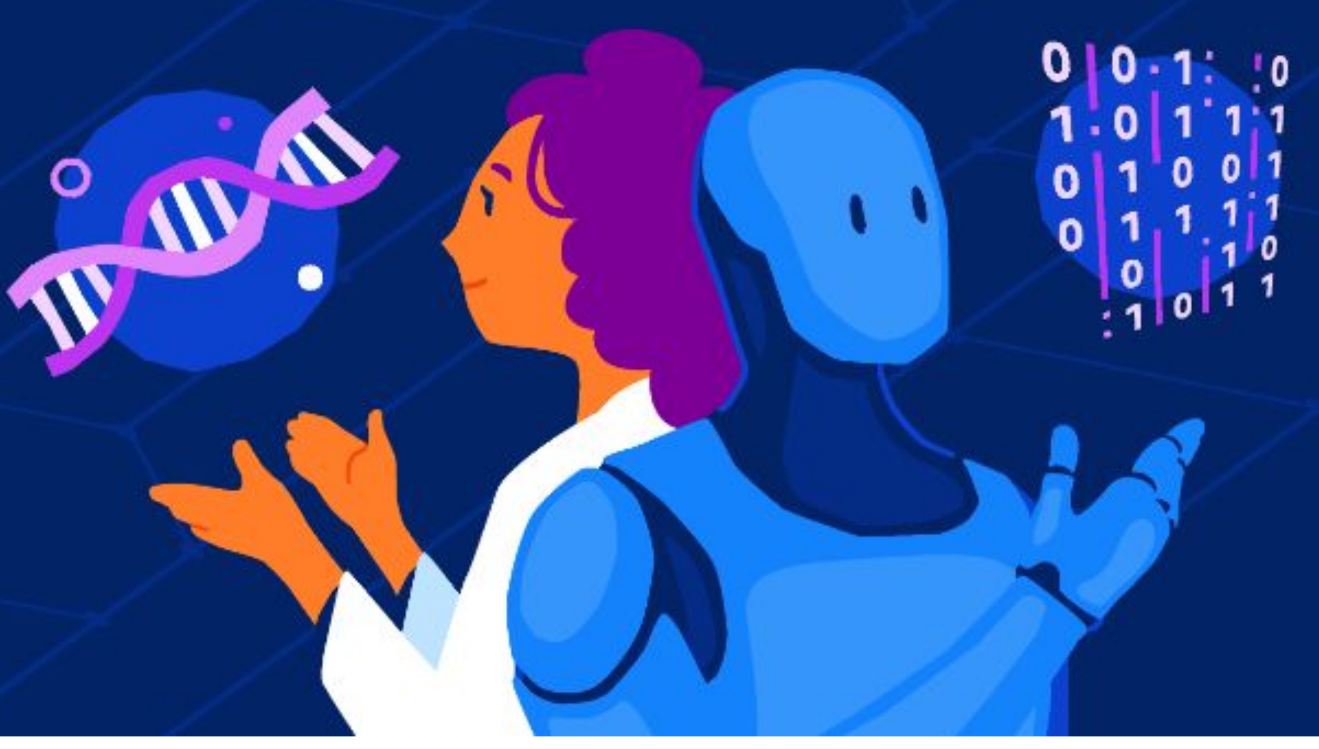
Roche UK

PP10

phuse

CSS 2025



Roche

Anonymisation of Unstructured Clinical Data: Patient Narratives as a Case Study

Nastazja Laskowski,
PhD Student in
Data Sciences
nastazja.laskowski@roche.com

Alex Hughes,
Principal Data Sciences
Product Leader
(industry supervisor)

Goran Nenadic,
Professor of Computer
Science at University of
Manchester (academic supervisor)

Mark Elliot,
Professor of Data Science
at University of Manchester
(primary academic supervisor)

Introduction

A challenging task at the intersection of several research areas

Anonymising unstructured clinical data combines natural language processing (NLP), to extract re-identifying or disclosive entities, with statistical disclosure control (SDC) to ensure participants cannot be re-identified. Together they ensure privacy while maintaining utility of data. This research evaluates barriers, facilitators and important considerations going forward.

Methods

Mixed-methods for a deep understanding of the challenge

This PhD research project in collaboration with Roche and University of Manchester adopts a mixed-methods approach, integrating interviews to capture expert insights, content analysis to uncover rich patterns in the data, and quantitative analysis to evaluate NLP performance and frequency of disclosive entities in Roche patient narratives from clinical study reports.

Results: 3 academic papers (pending publication)

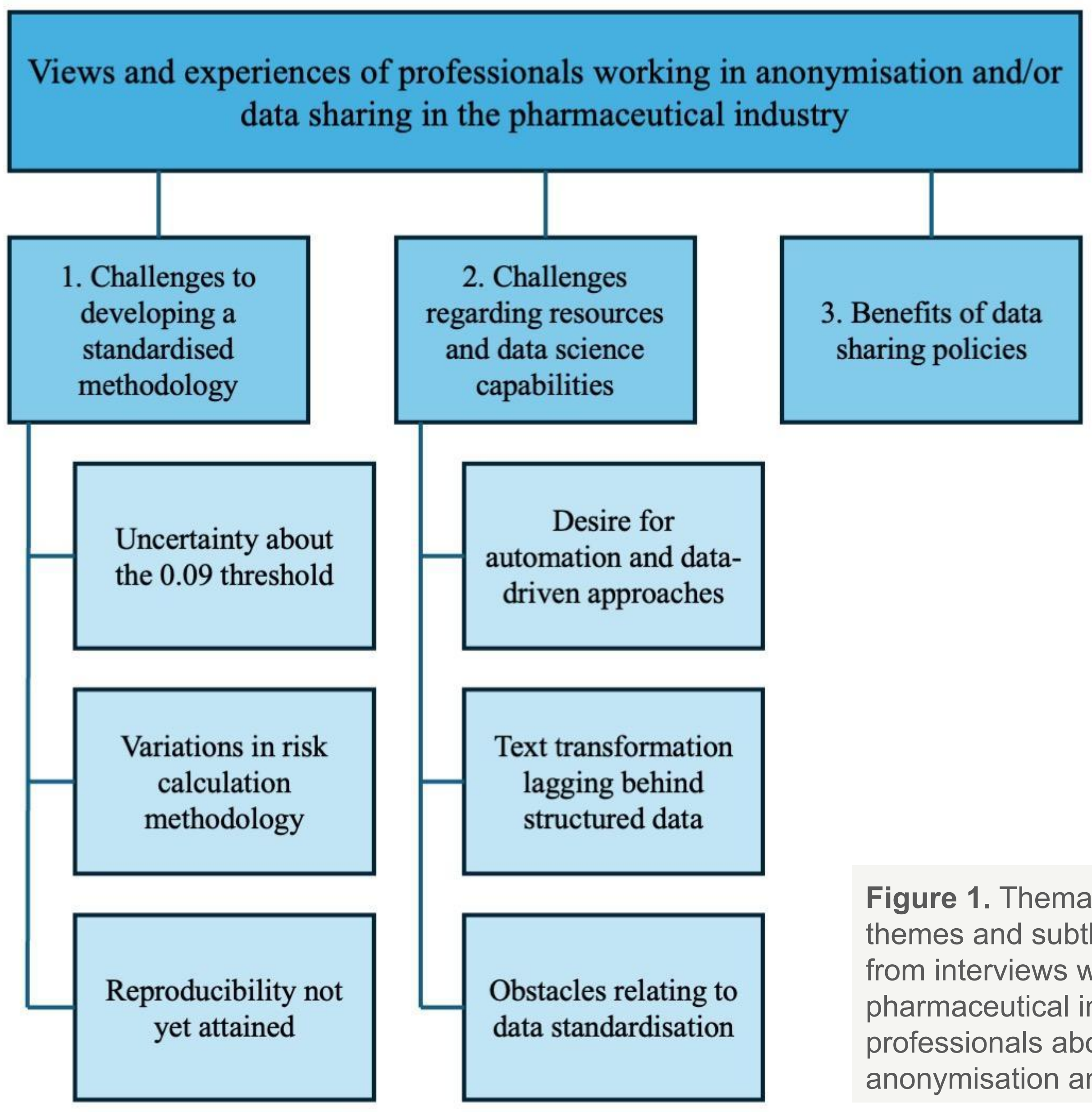
Research Question 1: **What are the barriers and facilitators to anonymisation according to pharma professionals?** *There needs to be flexibility in the risk of re-identification threshold, especially when anonymising small datasets, in order to preserve data utility. As of 2025, there is no cross-industry standardised 'gold-standard' methodology. Regulators are inconsistent in their acceptance and responses to methodologies. Automation and data-driven approaches are immature with text transformation lagging behind structured data (interview themes in Figure 1).*

Research Question 2: **Where do the challenges lie in relation to Natural Language Processing?** *Standard metrics for NLP performance do not translate well to the task of anonymisation. Different entities should have different importance (weights) when it comes to evaluating NLP performance. We find that the 'riskiness' of various entities is nuanced and dependent on the intruder and their knowledge-profile, although some are universally 'risky'. Certain therapeutic areas pose more challenges for NLP than others, such as psychiatry where meaning is often implied and standardised clinical language is not always used.*

Research Question 3: **How might generalisation of free-text terms work when applying SDC?** *MedDRA can be used to generalise adverse events (AEs) contained in free text patient narratives up a hierarchy. The benefits were that the hierarchies have a logical structure. However, there were multiple branches up which an AE could travel. Choices between branches need to be standardised as this impacts downstream processes.*

Conclusions

- Anonymisation of unstructured clinical data is challenging due to the varying risk of entities, and the semantic complexities in free text.
- MedDRA helps generalise adverse events but branching ambiguities in the hierarchy need standardisation to reliably meet SDC criteria.
- Increased knowledge sharing in the form of academic publications can help drive forward the convergence of NLP and SDC research to meet the needs of this field.



Views and experiences of professionals working in anonymisation and/or data sharing in the pharmaceutical industry

1. Challenges to developing a standardised methodology
- Uncertainty about the 0.09 threshold
- Variations in risk calculation methodology
- Reproducibility not yet attained

2. Challenges regarding resources and data science capabilities
- Desire for automation and data-driven approaches
- Text transformation lagging behind structured data
- Obstacles relating to data standardisation

3. Benefits of data sharing policies

Figure 1. Thematic map of themes and subthemes arising from interviews with pharmaceutical industry professionals about anonymisation and data sharing.