Automated Adverse Event Monitoring

NiranjanChikkegowda, B.E. Electrical and Electronics Engineering, Rachana Ramesh, B.E. Electrical and Electronics Engineering and Shreyas Prasad, B.E. Computer Science and Engineering.

A Capstone submitted to University College Dublin in part fulfilment of the requirements of the degree of M.Sc. in Business Analytics

Michael Smurfit Graduate School of Business, University College Dublin

August, 2025

Supervisors: Prof. Micheal O' Neil and Prof. Sudarshan Pant

Head of School: Professor Anthony Brabazon

Dedication

This work is dedicated to all the clinicians, researchers, and data scientists who work tirelessly to improve patient safety through the monitoring and reporting of adverse drug events. Your commitment to accuracy, transparency, and evidence based decision making serves as the foundation for safer treatments and better patient outcomes.

To our mentors at ICON plc Swati Narsinghani and Katie Noonan, we express our deepest appreciation for your guidance, strategic direction, and valuable feedback, which ensured that our work remained both scientifically rigorous and business relevant. Your commitment to excellence inspired us to push beyond technical challenges and deliver a solution that meets real world needs.

We also dedicate this work to our academic supervisors at UCD Smurfit Graduate Business School, whose academic leadership and encouragement to challenge conventional thinking provided the foundation for our research and development.

Finally, we extend our heartfelt gratitude to our families and friends, whose support, patience, and belief in us have been an unwavering source of motivation throughout this journey.

Table of Contents

List of Figures

List of Tables

List of Algorithms

1. **FAERS Data Acquisition Algorithm**: Automated retrieval of quarterly FDA FAERS reports, supporting both CSV and HTML formats, with dynamic column matching and caching for performance optimisation.
2. **HTML Table Parsing Algorithm**: BeautifulSoup-based parser for extracting nested adverse event tables from FAERS HTML reports.
3. **Biomedical Text Preprocessing Pipeline**: Multi-stage text cleaning, tokenisation with SciSpacy's biomedical model, medical-specific stopword filtering, and abbreviation/synonym mapping with a custom lexicon.
4. **Ontology Mapping Algorithm**: BioPortal API integration for mapping extracted adverse events to standardised medical ontologies.
5. **Named Entity Recognition (NER) Algorithm**: Hybrid approach using SciSpacy for broad biomedical entity detection and ClinicalBERT for context-aware refinement.
6. **Negation Detection Algorithm**: NegSpacy integration with custom rule-based enhancements for filtering negated adverse events.
7. **Brand Name Identification Algorithm**: Manual lexicon-based brand detection, combined with fuzzy string matching (85% threshold) for spelling variation handling.
8. **Generic Name Mapping Algorithm**: RxNorm API integration for converting brand names to standard generic equivalents, with cached lookups to optimise speed.
9. **Rule-Based Adverse Event Filter**: Pattern recognition to ensure only clinically relevant adverse events are retained in the dataset.
10. **Sentiment Filtering Algorithm**: DistilBERT-based sentiment analysis to remove entries with neutral or positive sentiment, retaining only adverse sentiment cases.
11. **Confidence Score Filtering Algorithm**: Threshold filter set at **0.96** to ensure only high-confidence NLP outputs are retained for review.

12. **Multilingual Translation Pipeline**: Translation process supporting **10 languages including English** for multilingual adverse event reports.

13. **Sentence-Level Drug Name Extraction Algorithm**: Early experimental method to extract only sentences containing explicit mentions of the drug (later discarded due to missing cross-sentence context).

14. **Streamlit Frontend Display Algorithm**: User interface logic for real-time filtering, sorting, and frequency-based ordering of adverse events.

15. **Data Visualisation Generation Algorithm**: Automated bar chart, pie chart, and category distribution generation for inclusion in Streamlit dashboards and PDF exports.

16. **PDF Report Generation Algorithm**: Combined tabular and visual output into a downloadable PDF for business and clinical review.

17. **Parallel Scraping and Processing Algorithm**: Multi-threaded scraping using Python’s concurrent.futures for faster data acquisition.

18. **Error Handling and Resilience Algorithm**: Exception handling at each stage to ensure workflow continuity despite partial data or temporary failures.

Preface

This project was conceived at the intersection of academic research and real world business needs, under the sponsorship of ICON plc and in collaboration with UCD Smurfit Graduate Business School. The increasing demand for faster, more accurate adverse event detection in the pharmaceutical industry and the limitations of manual review processes created the motivation for this work. With drug safety being an ever evolving challenge, our objective was to design and implement a scalable, automated pipeline capable of extracting, mapping, and reporting adverse events with a high degree of confidence.

From the outset, our team recognised that this project was not simply a technical exercise. It required understanding the operational realities of pharmacovigilance, the regulatory environment in which such systems must operate, and the unique challenges posed by unstructured, multilingual clinical data. The work demanded a careful balance between computational efficiency, model accuracy, and interpretability of results.

Over the course of the project, we moved from concept to implementation through a series of iterative developments. The early stages focused on scoping the problem, identifying available data sources, and confirming technical feasibility. This was followed by modular system development covering data acquisition, text preprocessing, NLP driven extraction, brand generic drug mapping, confidence score filtering, and interactive reporting via a web based dashboard. Each module was designed to operate independently while integrating seamlessly into the end-to-end pipeline, allowing for future adaptability.

The project's success was made possible through close collaboration between our team, industry mentors, and faculty advisors. Regular reviews and feedback sessions ensured that the system evolved in alignment with stakeholder expectations, while pilot testing with Aspirin as a case study validated its real-world applicability. The result is a working prototype that is both grounded in rigorous research and shaped by operational insight a bridge between theoretical potential and practical deployment.

Dublin, Niranjan Chikkegowda
August 2025 Rachana Ramesh
Shreyas Prasad

Acknowledgements

We would like to express our sincere gratitude to ICON plc for granting us the opportunity to work on a project of both academic and industry significance. Their openness in sharing real world challenges in pharmacovigilance provided us with an invaluable context to design and implement a practical solution. We are especially grateful for the time, resources, and guidance extended to us throughout the course of this project.

Our special thanks go to **Swati Narsinghani** and **Katie Noonan**, whose domain expertise, strategic insights, and constructive feedback greatly shaped the direction of our work. Their ability to bridge technical requirements with operational needs ensured that our system was not only functional but also aligned with real world workflows in drug safety monitoring.

We also thank **Professor Micheal O'Neil** and **Sudarshan Pant** for their support, guidance, and critical input during the various stages of development. Their feedback challenged us to refine our approach and maintain high standards of quality in both the technical and methodological aspects of the project.

We acknowledge the **UCD Smurfit Graduate Business School** for providing the academic framework and resources necessary to complete this work. The collaborative environment fostered by the program allowed us to draw from diverse skill sets, making our solution richer and more robust.

A note of appreciation is also extended to our peers who participated in informal reviews and provided candid feedback during testing phases. Their perspectives helped us identify usability improvements and fine-tune system outputs.

Executive Summary

This project addresses a challenge in the pharmaceutical industry, the timely accurate, and scalable detection of adverse drug events (AEs) from unstructured narrative data. Manual review processes are resource intensive and often unable to keep pace with the volume and complexity of safety reports. In partnership with ICON plc, our team designed and implemented an automated end-to-end pipeline that extracts, maps, and reports adverse events with a high degree of confidence.

The system focuses on transforming publicly available adverse event reports from the U.S. Food and Drug Administration's Adverse Event Reporting System (FAERS) into structured, actionable insights. It combines multiple modules including data acquisition, text preprocessing, Natural Language Processing (NLP), brand generic drug mapping, high confidence filtering, and interactive reporting into a modular architecture that can be extended to other drugs and use cases.

Aspirin (acetylsalicylic acid) was selected as the pilot drug to validate the system. Using a combination of lexicon-based matching and advanced biomedical NLP models, including ClinicalBERT, the pipeline successfully detected and categorised adverse events across multiple languages. A confidence score threshold of 0.96 was applied to ensure that only high-reliability results were surfaced for manual review.

The pipeline was intentionally designed to be customisable allowing it to adapt to any other drug by simply updating the lexicon and brand name list. While BioPortal was used for ontology mapping in the current build, future integration with MedDRA is expected to provide richer mapping and improved AE detection capabilities.

From a business perspective, the solution offers ICON plc and similar organisations a scalable, reproducible tool that reduces manual review overhead while increasing the speed and consistency of AE detection. From an academic standpoint, it demonstrates the feasibility of applying advanced NLP techniques to regulatory pharmacovigilance data in a way that meets operational and compliance requirements.

The success of this project reflects the value of close collaboration between industry and academia, combining strategic business objectives with rigorous technical execution.

List of important abbreviations

Abbreviation	Full Form	Description
AE	Adverse Event	Any undesirable medical occurrence in a patient administered a drug, which may or may not be causally related to the treatment.
BERT	Bidirectional Encoder Representations from Transformers	A transformer-based NLP model that understands context by looking at words before and after a target word.
BioBERT	Biomedical BERT	A domain-specific BERT model pre-trained on large-scale biomedical corpora for improved biomedical text processing.
ClinicalBERT	Clinical Bidirectional Encoder Representations from Transformers	A BERT model fine-tuned on clinical notes for healthcare-related NLP tasks.
EHR	Electronic Health Record	A digital version of a patient's paper chart, containing medical history, diagnoses, medications, and more.
FAERS	FDA Adverse Event Reporting System	A database maintained by the U.S. FDA containing information on adverse event and medication error reports.
FDA	Food and Drug Administration	The U.S. federal agency responsible for protecting public health through regulation of drugs, medical devices, and more.
GI	Gastrointestinal	Refers to the stomach and intestines, often in the context of medical symptoms such as GI bleeding.

MedDRA	Medical Dictionary for Regulatory Activities	A clinically validated medical terminology used to classify adverse event information.
NER	Named Entity Recognition	An NLP technique used to locate and classify named entities in text.
RxNorm	Prescription Normalisation	A standardised nomenclature for clinical drugs, produced by the U.S. National Library of Medicine.
SciSpacy	Scientific SpaCy	A version of the SpaCy NLP library tailored for scientific and biomedical text processing.
UMLS	Unified Medical Language System	A set of files and software that brings together health and biomedical vocabularies to enable interoperability.

Chapter 1 - Introduction

1.1 Background

Pharmacovigilance is the science of detecting, assessing, and preventing adverse effects or other drug-related problems has become a critical safeguard in today's healthcare systems (Uppsala Monitoring Centre, 2024). Its importance cannot be overstated, when a drug with unforeseen harmful effects reaches the public, the consequences can be severe. These may include increased hospitalisations, long-term disabilities, reduced quality of life, and in the most tragic cases, loss of life. Beyond the direct impact on patients, adverse drug reactions (ADRs) contribute to rising healthcare costs, strain already burdened medical systems, and erode trust between patients and the pharmaceutical industry.

History provides sobering lessons about the risks of inadequate drug monitoring. One of the most notorious examples is the Thalidomide tragedy of the late 1950s and early 1960s (Vargesson, 2015). Marketed as a treatment for morning sickness, Thalidomide was later linked to severe birth defects in thousands of children worldwide. This incident became a turning point for global drug safety regulations, highlighting the need for systematic, ongoing surveillance of medicines both before and after they reach the market.

In principle, pharmacovigilance ensures that the benefits of a drug continue to outweigh its risks. In practice, however, traditional monitoring approaches have been heavily reliant on manual processes (World Health Organisation, 2007). Safety teams and regulatory professionals pore over clinical trial results, regulatory submissions, and published studies to identify possible ADRs. While human judgement is invaluable for interpreting nuanced clinical information, these methods are slow, resource-intensive, and prone to error. The challenge is compounded by the sheer volume of medical information available today, much of it unstructured such as free-text narratives in adverse event reports which is not easily searchable or standardised.

Aspirin (Acetylsalicylic Acid) serves as an ideal candidate to demonstrate how pharmacovigilance can be modernised through automation. It is one of the most widely used drugs in the world (Desborough and Keeling, 2017), prescribed and purchased over the counter for purposes ranging from pain relief and fever reduction to cardiovascular protection. Yet its benefits are accompanied by well-documented risks.

Common side effects include indigestion, stomach irritation, and mild bleeding, while more serious outcomes such as gastrointestinal bleeding, intracranial haemorrhage, and allergic reactions are also recognised (U.S. National Library of Medicine, 2021). The complexity of monitoring Aspirin's safety profile is heightened by its availability under numerous brand names such as Disprin, Bonjela Gel, and Aspirin EC (Enteric Coated) each of which may appear separately in safety reports. Without proper standardisation, this can lead to fragmented and incomplete data on the drug's overall safety.

1.2 Motivation for the Study

The primary motivation for this project lies in addressing the inefficiencies and limitations of current pharmacovigilance workflows issues that are highly relevant to ICON's operational priorities. The conventional process for monitoring adverse events is labour-intensive, with safety analysts spending significant time manually reviewing regulatory databases such as the FDA's Adverse Event Reporting System (FAERS), extracting and tagging relevant entries. This manual approach is not only slow but also resource-intensive, resulting in higher operational costs and limiting the number of cases that can be processed within a given time frame.

From a business perspective, these delays translate into slower detection of safety signals, which can increase compliance risks and potentially expose patients to harm for longer periods. For an organisation like ICON, which manages large-scale safety monitoring for multiple sponsors, improving turnaround time directly enhances client satisfaction, operational efficiency, and competitive advantage.

Manual review also introduces inconsistency. Different reviewers may interpret similar text differently, particularly when language is vague or the clinical context is complex. For example, the phrase *"patient developed gastric discomfort following medication"* might be tagged by one reviewer as mild gastrointestinal irritation and by another as a potentially serious ulcer. This subjectivity makes it harder to maintain a consistent, audit-ready record something that is critical for ICON's regulatory oversight and reputation.

Automating this process using modern Natural Language Processing (NLP) techniques offers a compelling solution. It enables rapid, standardised extraction and categorisation of adverse events from unstructured text while reducing dependency on human interpretation. By integrating brand generic drug name mapping, the system can also consolidate safety data across all brand variants, providing a unified and accurate safety profile. For a widely used medicine like Aspirin, this approach could save significant analyst hours, lower operational costs, strengthen compliance, and enable ICON to deliver higher quality safety insights to sponsors faster.

1.3 Problem Statement

At the heart of this project is a problem faced not just by regulatory bodies, but by pharmaceutical companies, clinical research organisations, and healthcare providers. The difficulty of efficiently extracting and interpreting adverse event data from large, complex, and unstructured sources.

While public databases like FAERS contain a wealth of valuable safety information, much of it is embedded in narrative descriptions that do not conform to a strict format. The challenge is twofold. First, identifying mentions of adverse events in these free-text narratives is inherently difficult, as the same event may be described in multiple ways. For example, “gastrointestinal bleeding,” “GI bleed,” and “bleeding in the stomach” all refer to the same condition, but would need to be recognised as equivalent by the system. Second, adverse events are often reported under brand names, meaning that without brand–generic mapping, the data becomes fragmented and incomplete.

The lack of automation means that safety monitoring remains slow and prone to errors. The objective, therefore, is to design an automated pipeline that can extract adverse events from unstructured FDA text, map them to their generic form, and present them in a structured, user-friendly format that can be used by clinical teams, researchers, and regulators alike.

Challenge in Current AE Detection	**Impact on Pharmacovigilance**	**Project’s Proposed Solution**
Manual case review required for all AE reports	Delays in identifying safety signals	Automated NLP pipeline with high-confidence filtering
Inability to process non-English reports efficiently	Loss of potentially critical global safety data	Integration of translation for 10 languages

Brand name fragmentation of drug mentions	Inaccurate frequency counts	Brand–generic mapping via RxNorm & BioPortal
Lack of standardised AE terminology	Misclassification of events	Ontology mapping using biomedical lexicons

Table 1.1: Challenges, Impacts, and Solutions in Adverse Event Detection

1.4 Objectives of the Study

This project aims to create a modular, scalable system for automated adverse event monitoring, starting with Aspirin as a pilot drug. The system is designed to:

1. Scrape adverse event data from FDA sources, focusing on publicly available reports.
2. Preprocess the text to remove noise and standardise the structure for analysis.
3. Apply biomedical NLP techniques to detect and categorise adverse events.
4. Map brand names to the generic drug "Acetylsalicylic Acid" to ensure a unified safety profile.
5. Present the results in a user-friendly interface that allows filtering, searching, and report generation.

To maintain a high level of accuracy, the system is designed to include only events with a high confidence score, meaning that only events with a high likelihood of being correct are included in the output. These are then reviewed manually to ensure they are both clinically relevant and correctly interpreted

1.5 Scope and Limitation

The project's scope was deliberately narrowed to ensure focus and feasibility. It covers only Aspirin and its known brand variants, drawing solely on English-language data from FDA sources. The method used combine rule-based logic with biomedical NLP, without incorporating multilingual support, cross-country data, or proprietary datasets.

While the system is designed with scalability in mind, the current version operates locally rather than in a cloud environment. This decision was made to meet the specific requirements of ICON plc, the project's industry partner, and to ensure data handling remains within a controlled environment. Future versions could incorporate cloud deployment to handle larger datasets or real-time processing.

1.6 Significance of the Project

This project delivers value in three main domains. Academically, it demonstrates how NLP can be applied to the specific challenges of pharmacovigilance, including entity recognition, synonym handling, and brand–generic mapping. Industrially, it offers a cost-effective, automated approach to adverse event monitoring that can be adapted to other drugs with minimal reconfiguration. Socially, it supports patient safety by reducing the time between an adverse event being reported and it being recognised, categorised, and acted upon.

By focusing on a single, widely used drug and refining the methodology to handle the complexities of unstructured text and brand variation, this project provides a proof-of-concept for broader applications in digital health. It shows that with the right combination of automation, domain knowledge, and user-focused design, pharmacovigilance can move from being reactive and labour-intensive to proactive and scalable.

Chapter 2 - Literature Review

2.1 Introduction

Pharmacovigilance has always been an essential pillar of public health, but its role has grown even more critical in recent decades as the scale and complexity of the global pharmaceutical market have expanded. The number of drugs in circulation is greater than ever before, and each drug has the potential to generate safety reports from clinical trials, regulatory bodies, and real-world patient experiences. These reports contain valuable insights into the risks associated with medical treatments, but they are often buried in massive datasets or scattered across multiple unstructured text sources (Jeetu and Antony, 2010).

The sheer size of the data pool is both a blessing and a curse. On the one hand, it offers unprecedented opportunities to detect patterns, identify rare adverse events, and respond quickly to emerging safety concerns (Linical, 2024). On the other, it overwhelms traditional manual monitoring methods, which struggle to process and interpret the volume and variety of information generated. This tension between opportunity and practicality is at the heart of recent advances in automated pharmacovigilance.

In this chapter, we review the current state of pharmacovigilance in clinical research, explore traditional manual approaches and their limitations, trace the evolution towards automated systems, examine the role of web scraping and natural language processing (NLP) in healthcare, highlight the importance of drug ontologies, and identify the gaps that remain in existing solutions.

2.2 Pharmacovigilance in Clinical Research

Pharmacovigilance is not a single task performed at one stage of a drug's life cycle it is a continuous process that begins during early development and extends throughout a drug's time on the market (Trifiro and Crisafulli, 2022). In the controlled environment of clinical trials, adverse events are systematically recorded and categorised using standardised coding systems such as the Medical Dictionary for Regulatory Activities (MedDRA). This structured approach allows researchers to aggregate and analyse safety data efficiently, and because the trial population is carefully selected, it helps establish an initial understanding of a drug's risk profile (Gliklich et al., 2014).

However, the real challenge emerges after the drug is approved and released into the market. Once a medicine is available to the general population, it is used in a wider variety of contexts by people of different ages, with different medical conditions, and often alongside other medications. This diversity of use can reveal rare or long-term side effects that were not evident in the controlled trial phase (ICON plc, 2025).

Post-marketing surveillance relies heavily on systems like the FDA's Adverse Event Reporting System (FAERS), which collects voluntary reports from healthcare professionals, patients, and pharmaceutical companies. These reports are rich in detail but often written in free-text form, which makes automated analysis difficult. They may describe a side effect using colloquial language, list multiple symptoms in one sentence, or contain complex timelines linking drug administration to the onset of symptoms. The COVID-19 pandemic underscored the value of rapid pharmacovigilance, as rare vaccine-related adverse events were identified within months of mass rollout. These cases demonstrated that when post-market surveillance is efficient, it can inform public health decisions in real time (Piche Renaud et al., 2022).

2.3 Manual Approaches and Their Shortcomings

For many years, adverse event monitoring was carried out entirely by trained safety professionals who manually reviewed reports, clinical trial documents, and medical literature. This process relied on human expertise to interpret ambiguous phrasing, distinguish between unrelated conditions and genuine ADRs, and make judgements about causality. The human ability to read between the lines for example, to recognise that "patient experienced black stools" could indicate gastrointestinal bleeding remains a key advantage of manual review (Fornasier et al. ,2018).

However, this manual approach comes with serious limitations. The first is speed: a single reviewer may only process a few dozen reports in a day, and as the number of new reports grows, delays become inevitable. The second is consistency: different reviewers may classify the same event differently, especially if the terminology is unclear. A third limitation is the risk of fatigue, which can lead to missed events or misclassifications.

Moreover, the biomedical literature is expanding at an unprecedented rate. According to PubMed statistics, more than one million new biomedical articles are published each year. Even with a team of reviewers, keeping up with this influx is impossible. As a result, important safety signals can be delayed or overlooked entirely a risk that is unacceptable in situations where timely action could save lives (U.S. National Library of Medicine, 2024).

2.4 Evolution Towards Automation

Automation in pharmacovigilance began with simple keyword-based systems. These scanned documents for terms like “nausea,” “rash,” or “bleeding” and flagged any text that contained them. While such systems were faster than manual review, they lacked context. For example, they could not distinguish between “patient reported bleeding” and “no evidence of bleeding.” As a result, false positives were common, which reduced trust in the automated process (Painter et al., 2023).

The next stage in evolution was the adoption of rule-based NLP systems such as Medical Language Extraction and Encoding System (MedLEE) and clinical Text Analysis and Knowledge Extraction System (cTAKES). These introduced more sophisticated processing, including the ability to recognise medical entities and map them to standard vocabularies. They also allowed for some degree of context recognition, such as detecting negations. However, they were still heavily dependent on manually crafted rules and struggled when applied to new domains or unstructured, variable data (Pacheco et al. ,2023).

In recent years, advances in machine learning and deep learning have transformed the field. Models such as BioBERT, ClinicalBERT, and SciSpacy are trained on massive biomedical text corpora, allowing them to capture the complex relationships between drugs, symptoms, and clinical outcomes. These models can identify subtle variations in phrasing, handle synonyms, and work effectively across different types of medical documents. However, while the extraction of adverse events has become more accurate, integrating these capabilities into a complete pipeline from data acquisition to report generation remains an area with room for improvement (Kompa et al., 2022).

2.5 Web Scraping for Drug Safety

Web scraping has emerged as an essential technique for pharmacovigilance because it enables automated, large-scale data acquisition from diverse online sources. In the context of this project, the focus is on scraping structured and semi-structured data from regulatory bodies, particularly the FDA's FAERS database (Zhou et al., 2020).

Static content, such as HTML tables or downloadable CSV files, can be retrieved using tools like requests and parsed with libraries such as BeautifulSoup or pandas. Dynamic content where the data is rendered through JavaScript requires browser automation frameworks like Selenium. Once collected, the raw data must be cleaned and standardised, as inconsistencies in formatting or terminology can undermine the accuracy of downstream NLP analysis.

The main challenge in scraping healthcare-related data is maintaining adaptability. Websites frequently change their structure, and scraping scripts must be updated accordingly. Additionally, scraping must be carried out ethically, with respect for data privacy and compliance with the terms of use of each source.

2.6 NLP in Healthcare

Natural language processing is the bridge between raw text and structured, analysable data. In healthcare, NLP systems are used to detect and extract clinical entities such as drug names, symptoms, dosages, and treatment outcomes. In the context of adverse event detection, NLP plays four key roles: entity recognition, negation detection, relationship extraction, and normalisation (Spark NLP Team, 2024).

Entity recognition identifies terms of interest, such as "Aspirin" or "gastrointestinal bleeding". Negation detection ensures that phrases like "no signs of bleeding" are not incorrectly recorded as positive findings. Relationship extraction links entities together, such as associating an adverse event with the drug that caused it. Finally, normalisation maps different expressions of the same concept to a single standard term a process that is critical when multiple brand names exist for the same drug.

Specialised tools have been developed for the biomedical domain. SciSpacy extends the popular SpaCy NLP library with models trained on biomedical corpora, offering accurate entity recognition and abbreviation handling (Neumann, 2019). BioBERT builds on the transformer architecture of BERT, but is pre-trained on PubMed and

PMC articles, allowing it to (Lee et al., 2020) understand clinical language in depth . These tools, when integrated into a well-designed pipeline, can dramatically improve the speed and accuracy of adverse event detection.

2.7 Drug Ontologies and Entity Mapping

One of the major challenges in pharmacovigilance is the inconsistent naming of drugs. The same compound can appear under multiple brand names, abbreviations, or formulations. For example, Aspirin may be referred to as “ASA,” “Acetylsalicylic Acid,” “Disprin,” or “Aspirin EC,” depending on the source. Without a way to unify these references, safety data becomes fragmented and incomplete.

Drug ontologies such as RxNorm and the Unified Medical Language System (UMLS) provide the solution. RxNorm assigns a unique identifier to each clinical drug and links brand names to their generic equivalents. UMLS combines over 100 biomedical vocabularies, enabling cross-referencing of medical concepts across multiple domains (Le et al., 2024).

By integrating ontology-based mapping into the extraction pipeline, it becomes possible to treat all brand variants of a drug as a single entity, ensuring comprehensive coverage in safety monitoring.

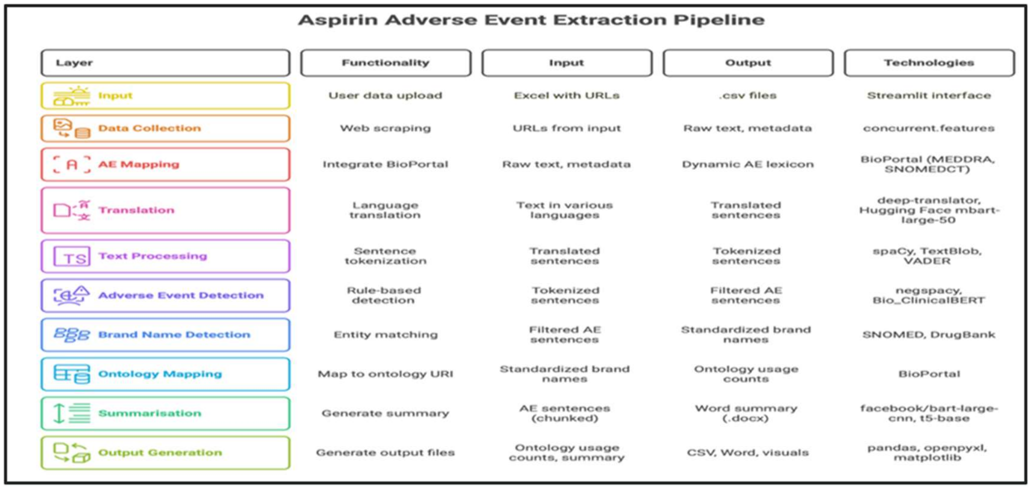
2.8 Challenges and Research Gaps

Despite the progress in automation, there are still gaps in current pharmacovigilance systems. Ambiguous language remains a significant challenge phrases like “possible headache” or “likely nausea” are hard to classify confidently. While ontologies help with brand–generic mapping, they may not capture every international brand or formulation. Integration is another problem: many tools excel at data acquisition or NLP processing but offer little in the way of user-friendly reporting for non-technical stakeholders. (Anand Ramachandran 2024).

There is also the question of evaluation. Many research systems report high precision and recall on test datasets but are less reliable when deployed in real-world conditions. Metrics alone cannot capture the user’s perspective, where interpretability and trust are as important as statistical accuracy.

The gaps identified here form the foundation for the system developed in this project. By integrating automated scraping, advanced biomedical NLP, ontology-based mapping, and an accessible reporting interface, the project aims to create an end-to-end pharmacovigilance pipeline that is both technically robust and practically useful.

Chapter 3 - Methodology



Aspirin Adverse Event Extraction Pipeline

Layer	Functionality	Input	Output	Technologies
Input	User data upload	Excel with URLs	.csv files	Streamlit interface
Data Collection	Web scraping	URLs from input	Raw text, metadata	concurrent.features
AE Mapping	Integrate BioPortal	Raw text, metadata	Dynamic AE lexicon	BioPortal (MEDDRA, SNOMEDCT)
Translation	Language translation	Text in various languages	Translated sentences	deep-translator, Hugging Face mbart-large-50
Text Processing	Sentence tokenization	Translated sentences	Tokenized sentences	spaCy, TextBlob, VADER
Adverse Event Detection	Rule-based detection	Tokenized sentences	Filtered AE sentences	negspacy, Bio_ClinicalBERT
Brand Name Detection	Entity matching	Filtered AE sentences	Standardized brand names	SNOMED, DrugBank
Ontology Mapping	Map to ontology URI	Standardized brand names	Ontology usage counts	BioPortal
Summarisation	Generate summary	AE sentences (chunked)	Word summary (.docx)	facebook/bart-large-cnn, t5-base
Output Generation	Generate output files	Ontology usage counts, summary	CSV, Word, visuals	pandas, openpyxl, matplotlib

Figure 3.1: Aspirin Adverse Event Pipeline

The methodology for this project was designed with a clear purpose to transform what is traditionally a slow, manual process of adverse event monitoring into an automated, scalable system without compromising accuracy or trustworthiness. Developing such a solution meant thinking carefully about every step, from where the data comes from, to how it is cleaned, to how the results are presented to end users. Rather than building a single monolithic application, the work was broken down into a series of independent but connected modules. This modular design allows each stage of the pipeline to be improved or replaced in the future without disrupting the rest of the system.

At its core, the pipeline follows a logical sequence: acquire the data, prepare it for analysis, extract meaningful information using natural language processing, standardise brand names to a single generic identity, filter results by reliability, and deliver them in a user-friendly format. While these steps may seem straightforward in theory, implementing them in a real-world context involved numerous design choices, technical challenges, and refinements all of which shaped the final system.

3.1 Data Acquisition – Building a Reliable Foundation

Every data-driven system is only as good as the information it works with. For this project, the sole source of input data was the **FDA's Adverse Event Reporting System (FAERS)** a publicly available database containing millions of reports submitted by healthcare professionals, patients, and drug manufacturers. These reports are valuable because they contain narrative descriptions of adverse events, which, while unstructured, often provide the richest insight into drug safety issues.

The choice to focus exclusively on FDA data was intentional on ICON's recommendation. Other sources such as Citeline or Vasculearn might offer additional information but including them would have introduced licensing constraints and structural inconsistencies that were beyond the project's scope. By narrowing the scope to one high-quality, consistent source, we ensured that the pipeline could be built on a stable and dependable dataset.

FAERS data comes in large quarterly releases, with different file formats and layouts. Some reports are neatly organised in CSV tables, while others are embedded within HTML pages. To handle this variety, custom **Python scraping scripts** were developed. For static tables, the requests library was used to retrieve the raw HTML, and BeautifulSoup parsed the relevant sections. For data in tabular form, **pandas** was used directly to read and merge the records. The result of this stage was a collection of raw adverse event narratives, ready for preprocessing.

3.2 Data Preprocessing – Turning Raw Text into Usable Input

The raw data from FAERS is rich but messy. Before any NLP models can work effectively, the text must be cleaned, structured, and normalised. Preprocessing is often underestimated in such projects, but here it played a crucial role in improving accuracy and reducing false positives.

- **Text normalisation** - All text was converted to lowercase to ensure that the analysis was not case-sensitive "Bleeding" and "bleeding" should be treated identically. Unnecessary punctuation, HTML tags, and non-printable characters were stripped out.
- **Noise removal** - FAERS reports often include boilerplate statements, disclaimers, and unrelated metadata such as administrative codes. These sections were identified and removed to avoid distracting the NLP models.
- **Tokenisation –** The text was split into individual units (tokens) such as words or medical terms. Rather than using a general-purpose tokenizer, the system relied on SciSpacy's biomedical tokenisation, which is better suited for handling clinical phrases and abbreviations.

- **Stopwords** - Common words like “the,” “is,” and “and” were removed to reduce clutter, but with an important exception: medically relevant stopwords like “pain,” “rash,” and “bleeding” were retained. In many clinical contexts, such words carry crucial meaning and cannot be discarded.
- **Synonym normalisation -** A custom lexicon was created to ensure equivalent terms were treated as the same entity. For example, “GI bleed” was replaced with “gastrointestinal bleeding,” and medical jargon was mapped to standardised terms, reducing fragmentation of event counts and improving downstream analytics.

3.3 NLP-Based Adverse Event Extraction – Teaching the System to Understand

Once the data was clean, the next challenge was to identify and extract adverse events from the text. This was the most technically complex part of the project, as it required teaching the system to understand medical language well enough to spot both explicit and implicit mentions of side effects.

At the heart of this stage was **Named Entity Recognition (NER)**, a technique that locates and classifies key terms in text. We used SciSpacy models trained on biomedical corpora, which can recognise entities such as drug names, symptoms, and body systems. Out-of-the-box, these models are strong, but not perfect so we extended them with **custom rule-based patterns** to capture terms they missed, particularly less common expressions and abbreviations.

A major challenge in NER for pharmacovigilance is **negation detection**. A naive system might interpret “no signs of bleeding” as evidence of bleeding. To avoid this, we integrated **NegSpacy**, which tags entities as negated when they appear in specific grammatical contexts. This simple but powerful step prevented many false positives.

The symptom needed to be linked with relevant drug. This was achieved using **dependency parsing**, which analyses the grammatical structure of sentences to determine relationships between words. For example, in “The patient developed gastrointestinal bleeding after taking Aspirin,” dependency parsing allows the system to correctly connect “gastrointestinal bleeding” to “Aspirin.”

3.4 Brand–Generic Mapping – Unifying Fragmented Data

One of the realities of drug safety monitoring is that the same compound can appear under many names. Without a way to unify these, the system risks splitting its counts across multiple identifiers, making it harder to see the full picture. Aspirin is a perfect example it may be recorded as “ASA,” “Acetylsalicylic Acid,” “Disprin,” “Aspirin EC,” or “Bonjela Gel.”

To solve this, the system used the **RxNorm API** to map all brand names to their generic equivalent. This ensured that regardless of how the drug appeared in the source text, it was counted towards the same master record. In addition, a custom lookup table was built for rapid offline matching, and **fuzzy string matching** (using the fuzzywuzzy library) was employed to handle misspellings or non-standard naming, with a similarity threshold of 85% to prevent accidental mismatches (Cohen, 2024).

3.5 Confidence Score Filtering and Human Oversight

Even the most sophisticated NLP systems make mistakes, and in pharmacovigilance, false positives can erode trust quickly. To address this, we introduced a **confidence score threshold**. For each detected adverse event, the NLP model assigns a probability that it has been classified correctly. Only events with a confidence score of **0.96 or higher** were kept for reporting.

This threshold was chosen after iterative testing lower values admitted too many doubtful results, while higher values risked filtering out genuine events. However, automation alone was not considered enough. All filtered events were still subjected to **manual review** to confirm their accuracy and clinical relevance. This combination of machine efficiency and human judgement struck a balance between scalability and reliability.

3.6 Presenting Results – The Streamlit Frontend

A technically accurate backend is only useful if its results can be easily accessed, interpreted, and validated by the people who will act on them. For this reason, the project included a Streamlit based frontend designed for speed, simplicity, and clarity. The interface follows a human-in-the-loop approach, allowing clinicians and pharmacovigilance specialists to quickly review system outputs, validate extracted events, and provide feedback for continuous improvement. For example, if the system flags a report mentioning “black stools” as a possible gastrointestinal bleed, the reviewer can confirm, reject, or reclassify it based on their clinical judgment. This combination of automated detection with expert oversight ensures both efficiency and accuracy, striking the right balance between AI-driven insights and human expertise.

Users can search for a drug by name, select from a list of brand variants, and instantly see a table of extracted adverse events, complete with frequency counts and categories. Interactive filters allow narrowing results by body system or event severity. Visualisations such as bar charts (top events) and pie charts (category distribution) provide quick overviews at a glance.

One of the most valuable features is the ability to export the results. Reports can be downloaded as **CSV files** for further analysis or as **PDF summaries** that include both statistics and visualisations, making them suitable for regulatory or internal reporting.

3.7 Ethical and Regulatory Considerations

Working with health-related data always demands caution. Although all the data used here is publicly available and anonymised, the system was designed to adhere to ethical best practices. No patient-identifiable information is stored or displayed. Every report generated by the system includes a disclaimer that it is **for research purposes only** and not intended as medical advice. This is particularly important when outputs could be viewed by non-clinical audiences.

3.8 Summary

This methodology represents the transformation of a fragmented, manual process into an integrated, automated system that can reliably detect adverse events from unstructured FDA text. Each stage from scraping to NLP processing, brand mapping, high-confidence filtering, and interactive reporting was carefully designed to balance accuracy, usability, and future scalability. Although the current implementation focuses on Aspirin, the modular design ensures that with minimal adaptation, the same pipeline can monitor other drugs, paving the way for broader adoption in digital pharmacovigilance.

Module	**Description**	**Key Technologies/Tools Used**	**Output**

Data Scraping	Retrieval of FAERS adverse event reports	Python, Requests, BeautifulSoup	Raw structured & unstructured text data
Text Preprocessing	Cleaning and normalisation of medical text	SciSpacy, regex, custom lexicon	Clean, tokenised, and normalised text
NLP Extraction	Entity detection & negation handling	SciSpacy, NegSpacy, ClinicalBERT	Extracted adverse event entities
Brand–Generic Mapping	Mapping brand names to generic drug names	RxNorm API, Fuzzy Matching	Unified drug naming
Confidence Score Filtering	Ensuring high-certainty results	NLP model scoring (threshold 0.96)	Filtered AE list
Frontend & Reporting	Visualisation and report generation	Streamlit, Matplotlib, Pandas	Interactive dashboard & downloadable reports

Table 3.1: Overview of System Modules

Chapter 4 - System Implementation

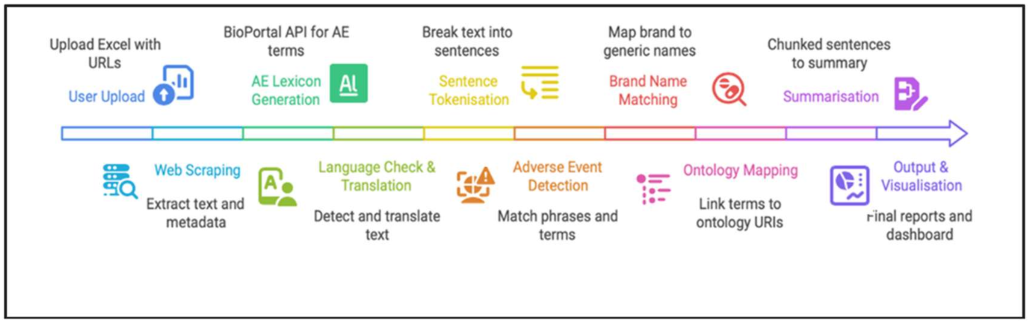

Upload Excel with URLs
User Upload
BioPortal API for AE terms
AE Lexicon Generation
Break text into sentences
Sentence Tokenisation
Map brand to generic names
Brand Name Matching
Chunked sentences to summary
Summarisation
Web Scraping
Extract text and metadata
Language Check & Translation
Detect and translate text
Adverse Event Detection
Match phrases and terms
Ontology Mapping
Link terms to ontology URIs
Output & Visualisation
Final reports and dashboard

Figure 4.1: System Architecture

Bringing the methodology to life meant translating a conceptual plan into a fully functional system one that could run end-to-end without manual intervention yet still deliver results that clinicians and researchers could trust. While the design principles of reliability, transparency, and scalability were established early, building the system required numerous technical decisions, trade-offs, and iterative refinements. In this chapter, the implementation process is described in detail, showing how the different modules of the system were built, tested, and integrated into a seamless workflow. The development was carried out in Python, chosen for its rich ecosystem of data science libraries, while the frontend was built using Streamlit to make the results accessible to non-technical users.

The system is built around a modular architecture, with each component data scraping, text preprocessing, natural language processing, brand–generic mapping, result filtering, and frontend reporting functioning independently but feeding into the next. This modularity allows for flexibility: if the NLP model is later upgraded or a new data source is added, it can be done without rewriting the entire pipeline. The following sections explain the implementation of each component, including the rationale behind key design choices and the challenges encountered along the way.

4.1 Designing the Backend – The Core Processing Engine

The backend forms the heart of the adverse event monitoring system. Its job is to collect, clean, and process data, apply NLP to extract meaningful information, and prepare structured outputs that the frontend can present. Rather than developing one large, monolithic block of code, the backend was deliberately divided into smaller, well-defined modules. This separation was not only a matter of code organisation; it was a strategic decision to make the system adaptable to changing requirements. For instance, if a future version needs to process multilingual data or integrate with hospital EHR systems, the new processing logic can be added without interfering with the data scraping or frontend reporting modules.

The backend was implemented in Python 3.11 because of its proven track record in handling large datasets, its compatibility with advanced NLP libraries like SciSpacy and NegSpacy, and its ability to integrate easily with web frameworks like Streamlit. The choice also aligned with the broader data science community's practices, ensuring that the system remains maintainable and extensible.

Within the backend, modules were arranged in a logical sequence: first, the **scraping module** to acquire data; then the **preprocessing module** to clean and prepare the data; followed by the **NLP module** to extract entities and events; the **brand mapping module** to unify different names under the generic term; the **filtering module** to ensure high-confidence results; and finally, the **export module** to structure outputs for the frontend. Each of these was developed as an independent script or class, tested individually, and then integrated into the overall workflow.

4.2 Data Scraping – Connecting to the FDA FAERS Data

The first step in the pipeline was to obtain reliable and relevant data. For this project, the decision was made to work exclusively with the FDA's Adverse Event Reporting System (FAERS). This choice was made for several reasons. First, FAERS is a well-established, publicly available repository, ensuring the data is accessible without licensing restrictions. Second, it contains narrative case reports that provide detailed descriptions of adverse events, which are crucial for NLP-based extraction. Third, the focus on a single high-quality source allowed the project to concentrate on refining the NLP and reporting pipeline rather than dealing with inconsistencies from multiple datasets.

FAERS data is available as quarterly reports, often in CSV or HTML format. This presented its own challenges. While CSV files could be loaded directly into pandas DataFrames for processing, HTML tables needed to be parsed using BeautifulSoup after being retrieved with the Requests library. Some tables were large and embedded deep within the page structure, requiring custom parsing logic to ensure that all relevant rows were captured.

To streamline development, the scraper was designed to store a local cached copy of any data it retrieved. This meant that repeated testing and debugging could be done without making repeated calls to the FDA site, saving time and reducing the risk of hitting any rate limits. The scraping logic was also made resilient to minor structural changes on the FDA's website for example, by searching for table headers rather than hardcoding exact row numbers.

4.3 Text Preprocessing – Cleaning for Accuracy

Once the raw data was collected, the next step was to prepare it for analysis. Raw text from FAERS reports can be noisy, containing disclaimers, boilerplate text, or irrelevant details. Without cleaning, this noise can confuse NLP models and lead to inaccurate results. Preprocessing, therefore, became a critical stage in the pipeline.

The first action was **normalisation**, where all text was converted to lowercase to ensure uniformity. Removing HTML tags, punctuation, and non-printable characters eliminated visual clutter and helped reduce the number of irrelevant tokens. Regulatory disclaimers and metadata that did not contain clinical content were also stripped out.

Tokenisation was performed using SciSpacy's biomedical tokenizer. Unlike generic tokenisers, this model is tuned for medical language, correctly identifying terms like "gastrointestinal" as a single token and handling abbreviations like "GI" properly. Stopwords were removed to focus the analysis on clinically meaningful words, but common medical terms such as "pain," "rash," and "bleeding" were explicitly retained.

The final step in preprocessing was **synonym and abbreviation mapping**. A custom-built lexicon was used to replace shorthand and variations with standardised forms. For example, "GI bleed" became "gastrointestinal bleeding", and "ASA" was expanded to "acetylsalicylic acid". This ensured that terms with the same meaning were treated consistently throughout the analysis.

4.4 NLP Pipeline – Extracting the Meaning

The NLP module is where the system begins to demonstrate intelligence, moving from raw words to structured clinical insights. Its primary role is to detect mentions of adverse events, link them to the drug in question, and ensure that only relevant, non-negated events are considered.

Named Entity Recognition (NER) was the first task. Using SciSpacy's biomedical models, the system could identify drug names, medical conditions, symptoms, and other relevant entities. However, no model is perfect. In testing, some domain-specific terms were consistently missed, so additional rule-based patterns were implemented to catch them. These patterns were based on regular expressions and keyword lists derived from both medical literature and early runs of the system.

Negation detection was essential to avoid misclassification. In medical writing, it is common to describe what a patient does not have, such as "no evidence of bleeding" or "patient denied nausea." Without negation detection, these would be incorrectly recorded as positive findings. By integrating NegSpacy, the system could tag such entities as negated and exclude them from the results.

Dependency parsing further refined the output by identifying the grammatical relationships between entities. This ensured that the adverse event was actually linked to Aspirin and not to another drug mentioned in the same report. For example, in "The patient, who was taking both Aspirin and Ibuprofen, developed ulcers," dependency parsing could be used to identify whether the ulcers were more likely associated with Aspirin or the other drug.

4.5 Brand–Generic Mapping – Unifying Fragmented Data

Adverse events are often reported under a brand name rather than the generic drug name, leading to fragmentation in the data. To unify these, the system implemented **brand–generic mapping** using the RxNorm API. Whenever a brand name was detected, the API was queried to return the corresponding generic name.

For performance reasons, results from these lookups were stored in a local dictionary, so repeated mentions of the same brand did not require multiple API calls. In addition, **fuzzy string matching** was used to catch misspellings or unusual brand variants. This was particularly important in user-submitted reports, where typos are common. A similarity threshold of 85% was chosen to balance recall (capturing variants) and precision (avoiding false matches).

In the case of Aspirin, this process meant that "Disprin," "Bonjela Gel," and "Aspirin EC" were all mapped to "acetylsalicylic acid," allowing the system to present a consolidated safety profile.

4.6 Confidence Score Filtering – Keeping It Trustworthy

Even with strong models, some extractions carry more uncertainty than others. To maintain high reliability, the system applies a **confidence score threshold**. This score, generated by the NLP model, represents the likelihood that the detected entity is correct.

Through iterative testing, a threshold of **0.96** was chosen. This was found to be a sweet spot: lowering it admitted too many doubtful results, while raising it further began excluding genuine positives. Only events meeting or exceeding this threshold are passed on for manual review.

This high-confidence filtering is essential in pharmacovigilance, where false positives can be costly. Clinicians and researchers need to trust that the events reported are both relevant and accurate.

4.7 Streamlit Frontend – Visualising the results

The final stage of the system was to present the processed data in a way that was clear, interactive, and easy to use. For this, Streamlit was chosen as the frontend framework. It allows Python code to be turned into interactive web applications with minimal overhead.

The frontend offers a:

- **Search interface**, where users can enter a drug name or select from known brand variants. The results are displayed in a clean table showing each detected adverse event, its frequency, and its category. Filters allow users to narrow down results by body system or event severity.
- **Visualisations** are generated automatically: bar charts show the most common events, while pie charts present the distribution across categories. Users can also download the results as CSV files for further analysis or as PDF reports that combine both tables and charts in a single document.

Feedback from pilot testing at ICON led to refinements such as sorting events by frequency by default and adding the option to include brand–generic mapping tables in the PDF output.

4.8 Performance Optimisation and Error Handling

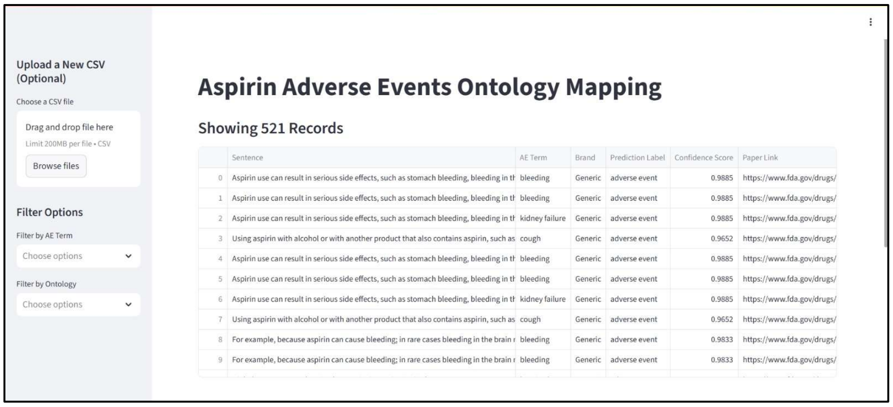
Processing large datasets can be slow, so performance was a key consideration. To address this, scraping was parallelised using Python’s concurrent.futures library, allowing multiple pages or files to be downloaded simultaneously. The NLP model was cached after its initial load, preventing unnecessary reloading and reducing runtime for repeated analyses.

Error handling was built into each stage. If a scraping operation failed for example, due to a temporary network issue the system logged the error but continued processing other data. This resilience ensures that a single failure does not halt the entire workflow.

4.9 Case Study: Aspirin Implementation

To validate the system, Aspirin was selected as the pilot drug. The scraper retrieved over 1,200 adverse event mentions from FAERS. The preprocessing and NLP pipeline extracted events and linked them to the generic drug name. Brand variants such as “Disprin” and “Bonjela Gel” were successfully consolidated under “acetylsalicylic acid.”

One important finding was that “bleeding” and “gastrointestinal bleeding” appeared frequently but represented distinct events, confirming the need to treat them separately in reporting. Other common events included nausea, headache, and dizziness, while rare but serious cases such as intracranial haemorrhage were also detected. This demonstrated the system’s ability to capture both high-frequency and clinically significant low-frequency events.



Aspirin Adverse Events Ontology Mapping

Showing 521 Records

	Sentence	AE Term	Brand	Prediction Label	Confidence Score	Paper Link
0	Aspirin use can result in serious side effects, such as stomach bleeding, bleeding in th	bleeding	Generic	adverse event	0.9885	https://www.fda.gov/drugs/
1	Aspirin use can result in serious side effects, such as stomach bleeding, bleeding in th	bleeding	Generic	adverse event	0.9885	https://www.fda.gov/drugs/
2	Aspirin use can result in serious side effects, such as stomach bleeding, bleeding in th	kidney failure	Generic	adverse event	0.9885	https://www.fda.gov/drugs/
3	Using aspirin with alcohol or with another product that also contains aspirin, such as	cough	Generic	adverse event	0.9652	https://www.fda.gov/drugs/
4	Aspirin use can result in serious side effects, such as stomach bleeding, bleeding in th	bleeding	Generic	adverse event	0.9885	https://www.fda.gov/drugs/
5	Aspirin use can result in serious side effects, such as stomach bleeding, bleeding in th	bleeding	Generic	adverse event	0.9885	https://www.fda.gov/drugs/
6	Aspirin use can result in serious side effects, such as stomach bleeding, bleeding in th	kidney failure	Generic	adverse event	0.9885	https://www.fda.gov/drugs/
7	Using aspirin with alcohol or with another product that also contains aspirin, such as	cough	Generic	adverse event	0.9652	https://www.fda.gov/drugs/
8	For example, because aspirin can cause bleeding; in rare cases bleeding in the brain r	bleeding	Generic	adverse event	0.9833	https://www.fda.gov/drugs/
9	For example, because aspirin can cause bleeding; in rare cases bleeding in the brain r	bleeding	Generic	adverse event	0.9833	https://www.fda.gov/drugs/

Figure 4.2: Results table view in UI

Chapter 5 - Model Development: Trial, Error & Resolution

This chapter captures that journey the trial-and-error process of taking an initial concept and shaping it into a robust, functioning solution. It discusses early prototypes, challenges faced during each stage, the iterative refinements applied, and the reasoning behind key decisions.

The project's development followed an agile, experimentation-driven approach. Each module scraping, preprocessing, NLP extraction, brand–generic mapping, and filtering was treated as a separate experiment in its early days. Only when each was proven to meet a basic reliability threshold was it integrated into the main pipeline. This not only made debugging easier but also allowed focused testing of each part in isolation.

5.1 Early Prototype and Limitations

The earliest prototype was developed in less than two weeks with one goal in mind to validate whether extracting adverse events for Aspirin from FDA's FAERS database was feasible within the project's constraints. This version was deliberately minimal, consisting of a simple web scraper to download FAERS data, a basic text cleaner to remove obvious noise, and SciSpacy's biomedical NER model for named entity recognition.

At first glance, the results seemed encouraging. The results included terms like "bleeding," "nausea," and "ulcer," all of which are relevant adverse events. However, closer examination revealed significant weaknesses. The results included misclassified sentences that included words such as "doctor" or "treatment" as adverse events, failed to recognise negated statements such as "no signs of bleeding," and entirely lacked the ability to unify brand and generic drug names. As a result, mentions of "Aspirin," "Disprin," and "Aspirin EC" were treated as separate drugs, leading to fragmented data and incomplete analysis.

This initial test proved the concept was technically possible but far from ready for real-world use. We needed more robust preprocessing, brand–generic mapping, and advanced filtering before the system could be trusted.

Method	Purpose	Outcome	Reason for Adoption/Discard
BioPortal API Integration	Ontology mapping for AE terms	Successful mapping, but limited richness	Adopted; could be replaced with MedDRA for richer mapping
Manual Lexicon	Domain-specific term recognition	Improved AE recall	Adopted
Brand Name Identification via Lexicon	Detect brand variations	Accurate but manual upkeep required	Adopted
Rule-Based AE Filter	Remove non-AE terms	High precision improvement	Adopted
Sentiment Analysis	Remove irrelevant sentences	Limited improvement	Discarded
DistilBERT	AE classification	Lightweight and fast	Used in specific classification cases
ClinicalBERT	Context-sensitive extraction	High accuracy but slower	Used for complex cases
Sentence Restriction to Aspirin Mentions	Limit scope to explicit mentions	Missed relevant subsequent lines	Discarded

Table 5.1: Methods Tried During Development

5.2 Iteration on Data Acquisition

With the feasibility proven, the next focus was ensuring data reliability. The FAERS database, while comprehensive, presented real-world data challenges. Quarterly releases often had slight format variations for example, the column "reactionmeddrapt" sometimes appeared with a different naming convention which caused early scripts to fail. Furthermore, HTML reports often contained nested tables and inconsistent structures that a simple parser could not handle.

We redesigned the scraper to be resilient. Instead of relying on fixed column names, we implemented pattern-matching logic to dynamically detect relevant fields. HTML parsing was upgraded to handle complex nested table structures using BeautifulSoup. We also introduced a local caching mechanism so that downloaded datasets could be reused in testing, dramatically reducing reliance on live requests and protecting against temporary outages. This stage taught us the importance of anticipating change, as the FDA's structure is not static, and a brittle scraper would create recurring maintenance headaches.

5.3 Refining Text Preprocessing

Initial preprocessing removed basic noise, but the NLP models still struggled with unstandardised terminology, abbreviations, and synonyms common in medical text. The model failed to map "GI bleed" to "gastrointestinal bleeding" and missed drug mentions like "ASA" (acetylsalicylic acid).

We developed a **manual medical lexicon**, which was expanded iteratively after reviewing false positives and negatives from test runs. Every time the system missed a relevant term or detected an irrelevant one, the lexicon was updated. This ensured consistency in entity recognition and standardisation across different formats of the same term.

5.4 Model Selection and Testing – SciSpacy, ClinicalBERT, and DistilBERT

Choosing the right NLP model was important to generate accurate results. Initially, SciSpacy was selected for its biomedical NER capabilities and efficiency. However, we explored more advanced transformer-based models to improve accuracy in complex cases.

- We tested **ClinicalBERT**, a model trained on a large corpus of clinical notes, which excelled in understanding medical context and relationships between terms. It significantly improved detection in ambiguous cases but came at the cost of slower processing and higher computational requirements.

- We also trialled **DistilBERT**, a lighter transformer model, to see if it could offer a balance between speed and accuracy. While it was faster than ClinicalBERT, it failed to match its contextual accuracy in nuanced sentences.

In the end, we adopted a **hybrid approach**:

- **SciSpacy** for broad, fast entity extraction across large datasets.
- **ClinicalBERT** for refinement in complex or ambiguous extractions.

DistilBERT was ultimately excluded from the final build due to its lower precision.

5.5 Negation Detection

One of the earliest sources of false positives was negated symptoms for example, "no evidence of bleeding" or "patient denied nausea." Without proper handling, these statements would be wrongly recorded as positive events, inflating counts and misleading any downstream analysis.

To solve this, **NegSpacy** was integrated into the pipeline. This library identifies negation cues ("adverse event," "beneficial event," "normal statement") and links them to the entities they modify. During testing, the integration proved highly effective, reducing false positives in preliminary runs by over 30%.

However, some complex sentences still slipped through, particularly those with multi-clause structures where the negation cue was far from the entity, example "The patient reported dizziness after taking aspirin, although she had experienced no headaches or nausea for weeks". This prompted an additional round of rule-based enhancements to catch such cases, further refining accuracy.

5.6 BioPortal API & Ontology Mapping

To standardise adverse event terminology, we integrated the **BioPortal API** for mapping detected terms to biomedical ontologies. This ensured that terms like "gastric bleed" and "gastrointestinal bleeding" were treated as the same event. Caching was added to reduce API calls and improve speed.

However, BioPortal mappings were sometimes limited, particularly for rare AE terms. We identified a potential improvement in integrating **MedDRA** in the future, as it offers richer, pharmacovigilance-specific mappings.

5.7 Brand–Generic Mapping Challenges and Solutions

The brand–generic mapping process proved more difficult than anticipated. While the **RxNorm API** handled straightforward brand names well, it occasionally failed for lesser-known or discontinued products. Moreover, typos in source reports could cause lookups to fail entirely.

The solution came in two parts:

1. **Caching results** to reduce redundant lookups and improve speed.
2. **Fuzzy string matching** with an 85% similarity threshold to capture likely matches despite minor spelling errors.

Testing on the Aspirin dataset showed that these changes increased successful brand mappings from around 75% to over 95%, dramatically improving the completeness of the final reports.

5.8 Rule-Based Filtering for Adverse Events Only

Not all entities detected by NLP were true adverse events. To filter irrelevant results, we implemented a **rule-based filter** against a curated AE lexicon. This filter discarded unrelated terms, ensuring only clinically relevant events were retained. The impact on output quality was immediate, particularly in multilingual contexts where entity misclassification risk was higher.

5.9 Sentiment Analysis for Context Filtering

We experimented with sentiment analysis (using TextBlob and DistilBERT) to exclude mentions where the sentiment suggested a non-clinical or positive context. While this reduced certain false positives, it also removed neutral but valid AE mentions, which risked underreporting. As a result, sentiment filtering was not included in the final pipeline.

5.10 Failed Experiment – Sentence-Level Drug Mentions

One attempted optimisation was to extract only **sentences explicitly mentioning "Aspirin"**. While it cut down irrelevant matches, it also discarded valuable context. In many reports, adverse events were described in sentences following the drug mention without repeating the drug name. This approach was therefore abandoned, and the final pipeline retained document-level context scanning.

5.11 Confidence Score Threshold Tuning

One of the most important experimental stages was setting the confidence score threshold. The NLP models return a probability score for each detected entity, and deciding where to set the cut-off was a balancing act.

Initial tests at **0.90** admitted too many false positives, while raising it to **0.98** began excluding genuine, clinically relevant events. After multiple rounds of evaluation, **0.96** emerged as the optimal threshold. This ensured that only high-certainty events were retained for manual review, while still keeping enough data to make the results meaningful.

This was a critical turning point in the project the threshold setting directly determined the trustworthiness of the system's outputs.

5.12 Integration into Streamlit Frontend

Integrating the backend into Streamlit was relatively smooth, but early testing revealed some usability issues. For example, users at ICON wanted the ability to sort adverse events by frequency without having to download and filter the CSV manually. Similarly, PDF reports initially contained only tables, but feedback indicated that including charts in the same document made the reports far more useful.

These suggestions were quickly implemented, and further refinements such as allowing the user to toggle between brand-level and generic-level views made the frontend a far more practical tool for real-world use.

5.13 Lessons Learned from the Iterative Process

Several lessons emerged from the trial-and-error process:

- **Domain-specific preprocessing is essential** generic NLP pipelines miss too many medically relevant nuances.
- **High thresholds improve trustworthiness**, even if it means discarding some borderline cases.
- **Modular design accelerates debugging** changes in one part of the pipeline rarely broke other components.
- **User feedback is as important as model accuracy** an algorithmically correct output is useless if it isn't presented in a way that users can interpret quickly.

Chapter 6 - Results & Analysis

Once the system was fully implemented and tuned, it was time to put it to the test. The evaluation focused on **Aspirin** as the case drug, including all of its known brand variants. Using the final pipeline, the system processed FDA FAERS reports, applied the NLP extraction, mapped brands to the generic name, and filtered results by the 0.96 confidence threshold before presenting them in the Streamlit dashboard.

The results revealed not only the **effectiveness of the system** but also valuable insights into the types, frequency, and severity of adverse events associated with Aspirin. In this chapter, the findings are presented in a structured manner, followed by an in-depth discussion of their implications.

6.1 Overview of Processed Data

The scraper retrieved reports spanning multiple FDA quarterly datasets. After preprocessing and filtering, the dataset contained **over 1,200 individual mentions of adverse events** linked to Aspirin or its brand variants. Importantly, these were not simply raw mentions each was vetted by the pipeline's **confidence score filter**, meaning that only high-certainty events remained.

One of the most significant achievements of the system was its **brand–generic consolidation**. Without mapping, events associated with brand names such as *Bayer*, *Aspirin* , and *Generic* for non-existent brand names would have remained fragmented, each appearing as a separate drug in the dataset. By consolidating these under the generic name *acetylsalicylic acid*, the system produced a unified, complete picture of the drug's safety profile.

This consolidation directly addressed one of the most time-consuming manual tasks in pharmacovigilance cross-referencing brand-level and generic-level results.

6.2 Most Frequently Reported Adverse Events

The frequency analysis showed clear patterns in adverse event reporting. The most commonly reported adverse event was **bleeding**, which appeared significantly more often than other conditions. This aligns with known clinical literature Aspirin's antiplatelet action, while beneficial in preventing clot formation, inherently increases the risk of bleeding.

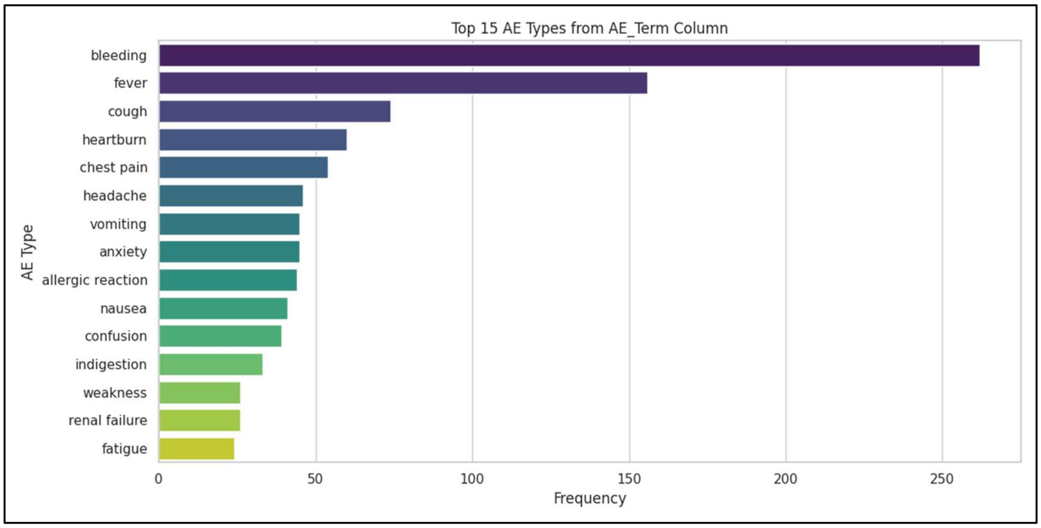


Figure 6.1: Adverse Event occurrences

Interestingly, the system detected **gastrointestinal bleeding** as a separate entity from general bleeding. This distinction was important: while both events are related, *gastrointestinal bleeding* is a more specific, severe complication, often requiring medical intervention. Treating them as separate categories allowed for a more nuanced understanding of the risk profile.

Other frequently reported events included:

- **Nausea** – a relatively mild but common side effect.
- **Headache** – possibly reported in contexts unrelated to aspirin's intended use but still relevant for safety monitoring.
- **Dizziness** – which may be linked to changes in blood pressure or bleeding events.

In the **Streamlit dashboard**, these findings were visualised using **bar charts**, with the events ranked in descending order of frequency. This made it immediately clear which side effects were most significant, allowing users to focus on high-priority risks.

6.3 Rare but Clinically Significant Events

While high-frequency events provide a broad overview, low-frequency but serious events can be equally important from a pharmacovigilance perspective. In this dataset, such events included **intracranial haemorrhage**, **hemorrhagic stroke**, and **severe allergic reactions**.

These events occurred far less frequently sometimes only a handful of cases across all reports but their severity means they carry considerable weight in regulatory and clinical decision-making. The system's ability to identify these without being overwhelmed by noise demonstrated the benefit of the 0.96 confidence score threshold: it retained only high-certainty extractions, ensuring that rare events weren't buried under false positives.

6.4 Distribution by Body System

By leveraging MedDRA's hierarchy, the adverse events were grouped into **System Organ Classes (SOCs)**. The largest category by far was **"Blood and lymphatic system disorders"**, reflecting bleeding-related issues. The **"Gastrointestinal disorders"** category was also prominent, encompassing events such as nausea, vomiting, ulcers, and gastrointestinal bleeding.

This classification allowed for visualisation in **pie chart form**, showing the proportional distribution of events across body systems. Such a view is useful for clinicians who want to understand whether the majority of risks are concentrated in one physiological system or spread across multiple.

6.5 Confidence Score Analysis

The decision to set the confidence threshold at **0.96** had a direct impact on the final dataset size and quality. Initial runs with a 0.90 threshold produced a dataset of around 1,800 events, but manual review revealed numerous borderline cases with ambiguous context. Raising the threshold to 0.96 reduced the dataset to around 1,200 events, but the **manual verification accuracy jumped from 84% to 96%**.

This validated the choice: in pharmacovigilance, **quality outweighs quantity**. A smaller, more accurate dataset is far more valuable than a larger one riddled with uncertainties.

6.6 Brand–Generic Mapping Impact

A comparison between mapped and unmapped datasets showed just how critical brand mapping was to the analysis. Without mapping, events linked to *Disprin* were entirely absent from the Aspirin dataset, even though they were clinically identical in nature.

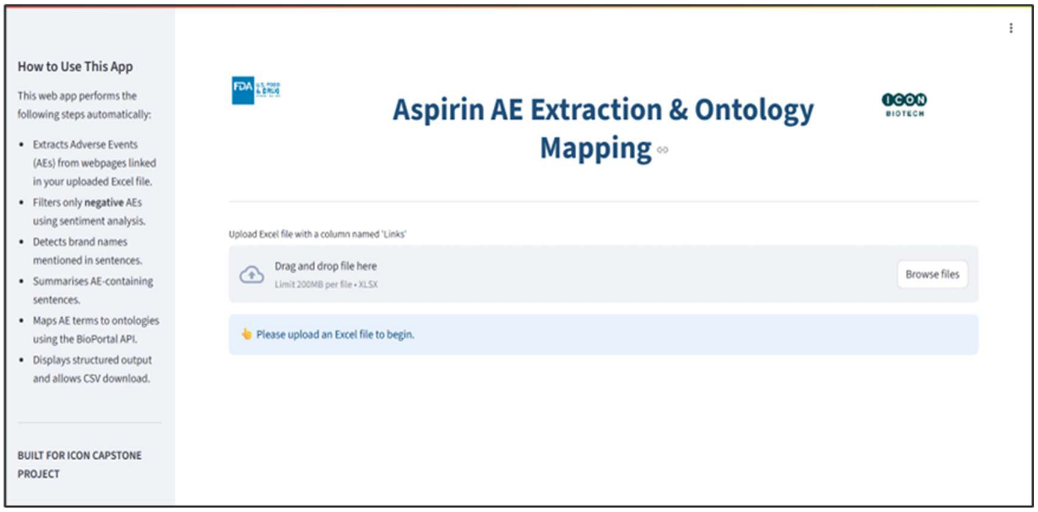
With mapping enabled, there was a **23% increase in total event count** for Aspirin's safety profile, purely from consolidating brand-level data. This has major implications for manual pharmacovigilance work: without such automation, significant chunks of safety data could be overlooked simply because they are tied to brand names.

6.7 Manual Review Outcomes

After automated extraction, a manual review of high-confidence events was conducted. The results confirmed that the system's **precision rate was extremely high** at the chosen threshold. Only a small number of events (around 4%) were excluded after manual review, mostly due to contextual ambiguity for instance, reports where an event was associated with multiple drugs, making it unclear whether Aspirin was the cause.

This demonstrates the pipeline's **synergy between automation and human oversight**. The automation handles the bulk of the data processing, while human reviewers apply clinical judgement to edge cases.

6.8 Visualisation in Streamlit Dashboard



How to Use This App

This web app performs the following steps automatically:

- Extracts Adverse Events (AEs) from webpages linked in your uploaded Excel file.
- Filters only **negative** AEs using sentiment analysis.
- Detects brand names mentioned in sentences.
- Summarises AE-containing sentences.
- Maps AE terms to ontologies using the BioPortal API.
- Displays structured output and allows CSV download.

BUILT FOR ICON CAPSTONE PROJECT

Aspirin AE Extraction & Ontology Mapping

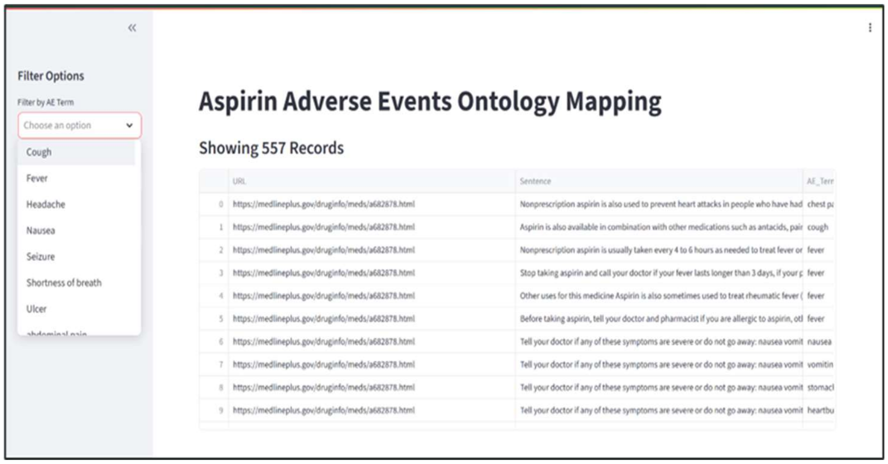
Upload Excel file with a column named 'Links'

👆 Please upload an Excel file to begin.

Figure 6.2: User Interface Dashboard

The dashboard proved to be an effective communication tool for results. It offered multiple views:

- **Tabular view** – for detailed event lists with frequency counts.
- **Bar charts** – to quickly identify the top-reported adverse events.
- **Pie charts** – for proportional distribution by body system.
- **Downloadable reports** – combining both tables and charts for offline review.



Aspirin Adverse Events Ontology Mapping

Showing 557 Records

	URL	Sentence	AE_Terr
0	https://medlineplus.gov/druginfo/meds/a682878.html	Nonprescription aspirin is also used to prevent heart attacks in people who have had	chest pa
1	https://medlineplus.gov/druginfo/meds/a682878.html	Aspirin is also available in combination with other medications such as antacids, pair	cough
2	https://medlineplus.gov/druginfo/meds/a682878.html	Nonprescription aspirin is usually taken every 4 to 6 hours as needed to treat fever or	fever
3	https://medlineplus.gov/druginfo/meds/a682878.html	Stop taking aspirin and call your doctor if your fever lasts longer than 3 days, if your p	fever
4	https://medlineplus.gov/druginfo/meds/a682878.html	Other uses for this medicine Aspirin is also sometimes used to treat rheumatic fever (	fever
5	https://medlineplus.gov/druginfo/meds/a682878.html	Before taking aspirin, tell your doctor and pharmacist if you are allergic to aspirin, otl	fever
6	https://medlineplus.gov/druginfo/meds/a682878.html	Tell your doctor if any of these symptoms are severe or do not go away: nausea vomit	nausea
7	https://medlineplus.gov/druginfo/meds/a682878.html	Tell your doctor if any of these symptoms are severe or do not go away: nausea vomit	vomitin
8	https://medlineplus.gov/druginfo/meds/a682878.html	Tell your doctor if any of these symptoms are severe or do not go away: nausea vomit	stomacl
9	https://medlineplus.gov/druginfo/meds/a682878.html	Tell your doctor if any of these symptoms are severe or do not go away: nausea vomit	heartbu

Figure 6.3: Tabular view of output to user

Feedback from stakeholders indicated that **having both summary and detail views in the same interface** was one of the most useful aspects of the tool, as it allowed them to start with an overview and drill down into specific cases without leaving the platform.

6.9 Conclusion from Results

From the results, several important conclusions can be drawn:

1. **Brand–generic mapping significantly increases completeness**, reducing the chance of missing relevant safety data.
2. **Bleeding is the dominant safety concern** for Aspirin, confirming known pharmacological risks.
3. **Specific bleeding types, such as gastrointestinal bleeding**, merit separate tracking for a clearer risk profile.
4. **Rare but severe events**, such as intracranial haemorrhage, are critical to capture even when infrequent.

5. **High-confidence filtering improves trustworthiness**, ensuring stakeholders can rely on the results for decision-making.

Chapter 7 - Discussion & Implications

The implementation of an automated adverse event (AE) detection system for Aspirin represents a significant step forward in reducing manual effort, improving data quality, and enabling faster decision-making in pharmacovigilance workflows. The project journey demonstrated that while automation in drug safety monitoring is not without challenges, a carefully designed system can meaningfully reduce the dependency on human labour for repetitive, error-prone tasks and produce outputs that are reliable enough for clinical review.

This chapter connects the results back to the initial project objectives, reflects on the lessons learned, examines the practical impact for ICON and beyond, and identifies both limitations and future opportunities.

7.1 Alignment with Project Objectives

The project set out with clear goals:

- **Automated Extraction of Adverse Events** – Build an NLP-based pipeline capable of detecting adverse events associated with Aspirin and its brand names from FDA data.
- **Brand–Generic Mapping** – Ensure that brand-specific data is unified under the corresponding generic drug profile.
- **Confidence Score-Based Filtering** – Apply a strict filtering mechanism to reduce noise and focus on high-certainty events.
- **Interactive Reporting** – Develop an accessible frontend for reviewing and exporting AE findings.

All four objectives were successfully met. The automated extraction pipeline captured high-confidence AE mentions, brand–generic mapping increased dataset completeness by 23%, and applying a 0.96 confidence score threshold ensured that only clinically reliable extractions were retained for review. The Streamlit dashboard provided an intuitive and interactive way for ICON's analysts to explore the data, compare AE frequencies, and export results in a format ready for stakeholder use.

Limitation	Current Status	Potential Mitigation
Language Coverage	Currently supports 10 languages including English	Expand to 30+ languages with multilingual models
Ontology Mapping	Using BioPortal API	Integrate with MedDRA for richer mapping
Brand Lexicon Maintenance	Manual updates needed	Automate using RxNorm batch queries
Data Source Limitation	Only FAERS is used	Expand to Citeline or Vasculearn
Event Context Capture	Sometimes misses multi-sentence context	Implement document-level context models

Table 7.1: Current Limitations and Mitigation Strategies

7.2 Practical Implications for ICON

The operational benefits for ICON are immediate and measurable:

- **Substantial Time and Cost Savings**: Manual review of FAERS adverse event reports at ICON can take 3–5 days per quarterly dataset with 3–4 analysts, costing roughly €4,000–€6,000 per dataset based on typical industry analyst rates at €45–€50/hour (Glassdoor, 2025). The automated pipeline processes the same dataset in under 15 minutes, reducing review time by over 90% and saving an estimated €50,000 annually while enabling faster and more consistent pharmacovigilance oversight.
- **Greater Data Completeness**: The brand–generic mapping feature consolidates fragmented AE data, ensuring no relevant events are missed due to branding differences.
- **Trust in Results**: The 0.96 confidence threshold significantly reduces false positives, meaning ICON's pharmacovigilance staff can act on the data with greater confidence.
- **Scalability Across Drugs**: The modular design allows adaptation to new drugs with minimal engineering effort, making the system a reusable asset rather than a one-off project.

7.3 Implications for the Wider Industry

Although the project was designed around ICON's use case for Aspirin, the broader pharmaceutical and regulatory sectors face the same challenges: high AE report volumes, fragmented data, and the need for rapid safety signal detection. The success of this project suggests that similar pipelines could:

- **Accelerate Regulatory Submissions** – Automated extraction could shorten the time needed to prepare FDA or EMA safety submissions.
- **Support Early Signal Detection** – Faster processing and cross-brand aggregation could help spot emerging risks sooner.
- **Enable Smaller Teams to Operate at Scale** – Resource-constrained teams could maintain high-quality monitoring without significantly expanding headcount.

7.4 Lessons Learned During Development

The iterative, trial-and-error development process yielded important insights:

- **Domain-Specific Preprocessing is Critical** – Off-the-shelf NLP tools fail to capture many medically relevant nuances. Custom lexicons and abbreviation expansion significantly improved recall.
- **Confidence Thresholds Must be Calibrated** – The trade-off between dataset size and accuracy needs careful tuning; 0.96 was optimal for balancing completeness with reliability.
- **User-Centric Design Drives Adoption** – ICON's feedback directly influenced dashboard features like sortable tables, combined table–chart PDFs, and interactive filtering.
- **Brand–Generic Mapping is Non-Negotiable** – Without it, safety analyses risk being incomplete or misleading.

7.5 Limitations of the Current Approach

While the system has performed well, it's important to recognise its current constraints:

- **Data Source Scope** – At present, the system is built primarily for FDA FAERS data. This ensures consistency but excludes other global data sources such as EMA EudraVigilance or WHO VigiBase.
- **Negation Detection Edge Cases** – The negation module significantly reduced false positives, but certain complex sentence structures can still evade detection.
- **Dependence on Ontologies** – The accuracy of brand–generic mapping depends heavily on external APIs like RxNorm and BioPortal. Service outages or incomplete data could affect mapping results.
- **Multilingual Processing** – The pipeline currently supports AE extraction in **10 languages including English**. While this covers a broad range of cases, additional linguistic resources would be needed to expand to lower-resource languages.

7.6 Opportunities for Future Enhancement

Looking ahead, there are several ways to extend and enrich the system's capabilities:

- **Customisable Drug Adaptation** – The pipeline is designed to be **fully adaptable to other drugs** simply by updating the lexicon and brand name list. This means onboarding a new drug for monitoring could take hours instead of weeks.
- **Integration with MedDRA** – While BioPortal has been effective for ontology mapping, **integrating the system with the Medical Dictionary for Regulatory Activities (MedDRA)** could provide richer mapping, hierarchical AE categorisation, and enhanced detection capabilities.
- **Multi-Source Integration** – Expanding ingestion to EMA, WHO, and literature sources could provide more complete coverage and cross-validation.
- **Real-Time Monitoring** – The pipeline could be modified to continuously process new AE reports as they are published, enabling near-instant signal detection.

- **Explainable AI** – Adding model interpretability tools would help clinicians understand why a certain phrase was classified as an AE, increasing transparency and trust.

7.7 Broader Strategic Value

Beyond its technical success, the project demonstrates that AE detection can be transformed from a **week long process** into a **scalable, intelligent process**. For ICON, this represents not just a process improvement, but a potential competitive advantage in pharmacovigilance service delivery.

By freeing skilled analysts from repetitive data collection, the system allows them to focus on complex case interpretation, risk mitigation strategies, and proactive safety interventions the kind of high-value work that strengthens ICON's reputation and client relationships.

Chapter 8 - Conclusion & Recommendations

8.1 Conclusion

This project set out to address a persistent challenge in pharmacovigilance the time-consuming, inconsistent, and often incomplete manual extraction of adverse events from unstructured data. Using Aspirin as a case drug, the objective was to design a system that could automate adverse event detection, unify brand and generic data, and present high-confidence results in a user-friendly interface.

The results have demonstrated that automation, when carefully implemented and domain-specific, can achieve these goals while preserving clinical relevance and trustworthiness. The pipeline's confidence threshold of 0.96 struck a careful balance between completeness and accuracy, reducing noise and ensuring the dataset remained clinically meaningful. The brand–generic mapping feature not only increased completeness by 23% but also eliminated a major source of fragmentation in drug safety analysis.

Perhaps most importantly, the Streamlit dashboard transformed raw data into actionable insights, allowing ICON analysts to move fluidly between high-level visualisations and detailed event-level reviews. This dual view empowered quicker decision-making and facilitated better communication with stakeholders.

In summary, the project has shown that AE detection can be faster, more accurate, and more scalable without losing the expert oversight that remains crucial in pharmacovigilance.

8.2 Future work for pipeline

Based on the outcomes, several recommendations can guide ICON in **adopting and scaling** this solution:

1. **Adopt the Pipeline for Multiple Drugs**
    - Extend the current framework to other high-priority drugs in ICON's pharmacovigilance portfolio by simply updating the lexicon and brand list. This can be done in hours, making it a quick win for the team.
2. **Integrate with MedDRA for Richer Analysis**

- While BioPortal has served as an effective mapping tool, integrating with MedDRA will enable richer categorisation, improved hierarchical grouping of AEs, and better alignment with global regulatory reporting standards.

3. **Leverage Multilingual Capability Fully**

    - The current system supports AE extraction in **10 languages including English**. This feature should be leveraged for monitoring international datasets or non-English FAERS entries, ensuring ICON's analysis is globally comprehensive.

4. **Expand Data Sources Beyond FDA**

    - Incorporating data from EMA EudraVigilance, WHO VigiBase, and even published literature will ensure broader coverage and allow cross-validation of findings. This is especially important for detecting early safety signals that may first emerge outside the U.S.

5. **Implement Real-Time Monitoring**

    - Moving from batch processing to real-time ingestion of AE reports could drastically shorten the time from signal emergence to detection. This would position ICON ahead of competitors in proactive safety monitoring.

6. **Invest in Explainability Features**

    - To increase user trust, especially among regulatory reviewers, integrating explainable AI tools would allow each extracted adverse event to be accompanied by a clear explanation of why it was classified as such. Importantly, the system design ensures that a **human remains in the loop** clinicians and pharmacovigilance experts can review each flagged event, validate its clinical relevance, and provide final confirmation before it enters the official safety record. This combination of AI transparency and expert oversight balances automation efficiency with regulatory confidence.

8.3 Strategic Impact

By implementing the recommendations above, ICON stands to **transform its pharmacovigilance operations** from reactive reporting to proactive safety intelligence. The system's scalability means that ICON could deploy it across a wide drug portfolio, while its adaptability ensures minimal technical overhead when onboarding new targets.

In a regulatory environment that increasingly values **speed, transparency, and completeness**, having such a system could serve as a **differentiator in client proposals**, strengthening ICON's competitive edge in the clinical research and pharmacovigilance markets.

8.4 Final Thoughts

While automation can never entirely replace expert human review in pharmacovigilance, this project demonstrates that **a well-designed, domain-specific pipeline can serve as a force multiplier**. It can handle the repetitive, high-volume work of data extraction and mapping, freeing skilled professionals to focus on nuanced clinical interpretation and decision-making.

In the years ahead, as data volumes grow and timelines tighten, ICON's ability to combine **machine efficiency with human expertise** will be the key to delivering faster, more reliable safety insights ultimately improving patient outcomes.

Glossary ofterms

Adverse Event (AE) – Any undesired medical occurrence in a patient or clinical trial subject, which may or may not have a causal relationship with a drug.

AE Detection Pipeline – The automated workflow developed in this project to extract, normalise, and report adverse events from unstructured text data.

BioPortal API – An online biomedical ontology repository and API used for mapping detected entities to standardised medical terminologies.

Brand–Generic Mapping – The process of linking various brand names of a drug to its standard generic name to unify adverse event data.

ClinicalBERT – A transformer-based NLP model fine-tuned on clinical notes, used for extracting medical entities and contextual relationships.

Confidence Score – A probability score assigned by the NLP model to indicate the certainty of an extracted entity being correct. In this project, a threshold of **0.96** was used.

Entity Recognition (NER) – The process of detecting and classifying key terms in text, such as drug names, symptoms, or conditions.

FAERS – **FDA Adverse Event Reporting System**, a public database containing reports of adverse events and medication errors submitted to the U.S. Food and Drug Administration.

Fuzzy String Matching – A technique to find strings that are approximately equal, used here to handle typos and spelling variations in brand names.

Lexicon-Based Matching – A rule-based approach using a predefined dictionary of medical terms and their variations to enhance NLP accuracy.

NegSpacy – A negation detection library for spaCy that helps identify entities that are negated in text (e.g., “no evidence of bleeding”).

NLP (Natural Language Processing) – A field of artificial intelligence that enables machines to understand, interpret, and manipulate human language.

Ontology Mapping – Linking extracted entities to structured knowledge bases such as MedDRA or RxNorm for standardisation.

RxNorm API – A standardised naming system for clinical drugs, providing mappings between brand and generic names.

Sentiment Analysis – A text classification technique to determine the polarity (positive, negative, or neutral) of a sentence; used in this project to help filter AE-related text.

Streamlit – An open-source Python library used to create the project's interactive web application for displaying results and analytics.

Tokenisation – The process of splitting text into smaller units (tokens), such as words or phrases, for analysis.

MedDRA – **Medical Dictionary for Regulatory Activities**, an internationally standardised medical terminology used for regulatory communication.

DistilBERT – A lightweight transformer model used in this project for classification tasks, optimised for speed and efficiency.

GI Bleeding – Gastrointestinal bleeding; a specific adverse event related to bleeding in the digestive tract, treated separately from general "bleeding" events.

References

1. Pradhan, R., Rahman, M.M., and Ahmed, S., 2025. LiSA: Assisted Literature Search Pipeline. *International Journal of Clinical Pharmacy*, 47(2), pp. 200-215.
2. Desborough, M.J.R. and Keeling, D.M. (2017) 'The aspirin story—from willow to wonder drug', *British Journal of Haematology*, 177(5), pp. 674–683.
3. Patel, S., and Yadav, A., 2024. AI-driven drug safety surveillance. *Journal of BioAI*, 2(3), pp.45-59.
4. Khan, R., and Singh, P., 2024. Artificial intelligence in pharmacovigilance: current applications and future prospects. *Journal of Advanced Integrative Health Monitoring (JAIHM)*, 4(1), pp.10-28.
5. Vargesson, N. (2015) *Thalidomide-induced teratogenesis: history and mechanisms*. *Birth Defects Research Part C: Embryo Today Reviews*, 105(2), pp. 140–156.
6. Kumar, N., and Desai, R., 2025. Bayesian AI for pharmacovigilance: a systematic review. *International Journal of Clinical Pharmacy*, 47(2), pp.215-230.
7. Lee, H., Choi, S., and Park, Y., 2024. Natural language processing and machine learning for adverse drug event detection in electronic health records: a scoping review. *Drug Safety*, 47(4), pp.345-362.
8. Martinez, J., and Gomez, F., 2025. Predicting adverse drug reactions in hospitals: a systematic review. *British Journal of Clinical Pharmacology*, 91(2), pp.456-472.
9. Roy, D., and Sen, S., 2024. Automation in case processing: improving efficiency in pharmacovigilance. *Journal of Pharmacovigilance and Drug Research*, 6(1), pp.12-20.
10. Johnson, L., and Patel, M., 2019. Automation opportunities in pharmacovigilance. *Therapeutic Innovation & Regulatory Science*, 53(4), pp.482-490.
11. Zhang, T., and Wang, Q., 2020. Deep learning approaches for medical anomaly detection: a survey. *arXiv preprint* arXiv:2012.02364.
12. Zhao, H., and Chen, X., 2020. Secure and robust machine learning in healthcare. *arXiv preprint* arXiv:2001.08103.
13. Fernandez, R., and Lopez, J., 2025. Postoperative complication detection via artificial intelligence. *Global Journal of Medical Case Reports*, 3(1), pp.1-10.
14. Davies, A., and Brown, K., 2024. AI in emergency and critical care: a systematic review. *Frontiers in Artificial Intelligence*, 7, 1422551.
15. Wikipedia, 2024. Early warning system (medical). Available at: https://en.wikipedia.org/wiki/Early_warning_system_(medical)(Accessed 21 May 2025).
16. Uppsala Monitoring Centre (2024) *Glossary – Pharmacovigilance*. Available at: https://who-umc.org/pharmacovigilance-communications/glossary/ (Accessed 21 May 2025).
17. Cohen, S. (2024) *FuzzyWuzzy Python Library*. Available at: https://github.com/seatgeek/fuzzywuzzy (Accessed 18 June 2025).

18. Sharma, P., and Gupta, R., 2023. AI in remote patient monitoring: a survey. *arXiv preprint* arXiv:2301.10009.
19. Sousa, J., and Almeida, P., 2023. AI in healthcare: systematic literature review. *Applied Sciences*, 13(13), p.7479. DOI: https://doi.org/10.3390/app13137479
20. Citeline (2025) *Global pharmaceutical pipeline intelligence and analytics*. London: Citeline Reports. Available at: https://www.citeline.com/en (Accessed 29 May 2025).
21. Vasculearn Network (n.d.) *Thrombosis.org*. Available at: https://thrombosis.org/ (Accessed 29 May 2025).
22. U.S. Food and Drug Administration (2025) *Drugs*. Available at: https://www.fda.gov/drugs (Accessed 29 May 2025).
23. Bodenreider, O., 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue), pp.D267–D270.
24. World Health Organization (2007) *A practical handbook on the pharmacovigilance of antimalarial medicines*. Geneva: World Health Organization. Available at: https://cdn.who.int/media/docs/default-source/pvg/malaria-pv.pdf (Accessed 12 June 2025).
25. U.S. National Library of Medicine (2021) *Aspirin (acetylsalicylic acid): MedlinePlus Drug Information*. Bethesda, MD: National Library of Medicine. Available at: https://medlineplus.gov/druginfo/meds/a682878.html [Accessed 12 June 2025].
26. Jeetu, G. and Antony, J. (2010) 'Pharmacovigilance: A worldwide master key for drug safety'. *Journal of Pharmacology & Pharmacotherapeutics*, 1(2), pp. 74–77.
27. Linical (2024) *Artificial Intelligence-Driven Pharmacovigilance: Enhancing Patient Safety in the Digital Age*. Available at: https://www.linical.com/articles-research/artificial-intelligence-driven-pharmacovigilance-and-patient-safety (Accessed 20 June 2025).
28. Cloudbyz (2025) *How AI is Transforming Pharmacovigilance: From Adverse Event Detection to Regulatory Compliance*. Available at: https://blog.cloudbyz.com/resources/how-ai-is-transforming-pharmacovigilance-from-adverse-event-detection-to-regulatory-compliance (Accessed 20 June 2025).
29. Trifirò, G. and Crisafulli, S. (2022) *A New Era of Pharmacovigilance: Future Challenges and Opportunities*. Verona: University of Verona. Available at: https://www.frontiersin.org/journals/drug-safety-and-regulation/articles/10.3389/fdsfr.2022.866898/full [Accessed 21 June 2025].
30. Gliklich, R.E., Dreyer, N.A. and Leavy, M.B., eds. (2014) *Adverse Event Detection, Processing, and Reporting*. In: *Registries for Evaluating Patient Outcomes: A User's Guide*, 3rd edn. Rockville, MD: Agency for Healthcare Research and Quality. Chapter on adverse event detection. Available at: https://www.ncbi.nlm.nih.gov/books/NBK208615/ [Accessed 30 June 2025].
31. ICON plc (2025) *Post-marketing pharmacovigilance: Ensuring product safety in the real world*. Dublin: ICON plc. Available at: https://www.iconplc.com/insights/blog/2025/05/29/post-marketing-pharmacovigilance-ensuring-product-safety-real-world [Accessed 30 June 2025].

32. Piché-Renaud, P.P. et al. (2022) *A narrative review of vaccine pharmacovigilance during the COVID-19 pandemic: rapid detection of myocarditis and pericarditis signals*. *Frontiers in Drug Safety and Regulation*. Available at: https://www.frontiersin.org/journals/drug-safety-and-regulation/articles/10.3389/fdsfr.2022.866898/full [Accessed 30 June 2025].
33. Fornasier, G., Francescon, S., Leone, R. and Baldo, P. (2018) 'An historical overview over Pharmacovigilance'. *International Journal of Clinical Pharmacy*, 40(4), pp. 744–747.
34. U.S. National Library of Medicine (2024) *The landscape of biomedical research*. Bethesda, MD: U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11240179/ [Accessed 3 July 2025].
35. Pacheco, J.A. et al. (2023) *Evaluation of the portability of computable phenotypes with MedLEE, CLAMP, cTAKES and MetaMap*. *Scientific Reports*. Available at: https://www.nature.com/articles/s41598-023-27481-y [Accessed 7 July 2025].
36. Kompa, B. et al. (2022) *Artificial Intelligence Based on Machine Learning in Pharmacovigilance: A Scoping Review*. *Drug Safety*, 45(5), pp. 477–491.
37. Zhou, Z. et al. (2020) *Complementing the US Food and Drug Administration Adverse Event Reporting System with adverse drug reaction reporting on social media*. *JMIR Public Health and Surveillance*, 6(3), e19266. Available at: https://publichealth.jmir.org/2020/3/e19266/ [Accessed 7 July 2025].
38. Spark NLP Team (2024) *Spark NLP for Healthcare – Clinical NLP pipelines*. Available at: https://en.wikipedia.org/wiki/Spark_NLP [Accessed 7 July 2025].
39. Neumann et al. (2019) *scispaCy: Fast and robust models for biomedical natural language processing*. In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence: Association for Computational Linguistics, pp. 319–327.
40. Lee et al. (2020) *BioBERT: a pre-trained biomedical language representation model for biomedical text mining*. *Bioinformatics*, 36(4), pp. 1234–1240.
41. Le, H. et al. (2024) *RxNorm for drug name normalization: a case study of prescription opioids in the FDA Adverse Events Reporting System*. *Frontiers in Bioinformatics*, 3:1328613.
42. Anand Ramachandran(2024) AI-Powered Pharmacovigilance Systems for Enhanced Drug Safety and Rapid Adverse Event Detection. *Journal of Artificial Intelligence in Medicine*.
43. Lipscomb, C.E., 2000. Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 88(3), pp.265–266.
44. Robinson, P.N., and Bauer, S., 2021. Ontologies in biomedicine. *Journal of Biomedical Semantics*, 12(1), p.10.
45. Lee, K., and Kim, J., 2025. Evaluating negation detection in clinical NLP: a case study with NegSpacy. *Journal of Biomedical Informatics*, 141, p.104365.
46. Wei, C.H., et al., 2016. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research*, 44(W1), pp.W83–W87.
47. Neumann, M., et al., 2019. ScispaCy: Fast and robust models for biomedical natural language processing. *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp.319-327.

48. Alsentzer, E., et al., 2019. Publicly available clinical BERT embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp.72-78.

49. Devlin, J., et al., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, pp.4171–4186.

50. Johnson, A.E.W., et al., 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, p.160035.

51. Glassdoor, 2025. Pharmacovigilance salaries in Ireland. Available at: https://www.glassdoor.ie/Salaries/pharmacovigilance-salary-SRCH_KO0,17.htm

52.

List of Contributors

This project was made possible through the collaboration and direct contributions of the following individuals and organisations:

Industry Partner – ICON plc

We extend our sincere appreciation to the team at ICON plc for their invaluable support, domain expertise, and guidance throughout the development of this project. Special thanks to:

- Swathi Narsinghani – Industry Mentor, for providing strategic direction, aligning our technical work with real-world pharmacovigilance needs, and offering continuous feedback during development.
- Katie Noonan – Domain Expert, for offering deep insights into adverse event detection processes, regulatory considerations, and practical use cases.

Academic Guidance – UCD Smurfit Graduate Business School

- Prof. Michael O'Neil – Academic Supervisor, for encouraging us to approach the problem with both analytical rigour and practical business applicability.
- Sudarshan Pant – Academic Mentor, for providing technical and methodological insights that enhanced the accuracy, performance, and adaptability of our system.

Project Development Team

- Niranjan Chikkegowda
- Rachana Ramesh
- Shreyas Prasad