MSR129

# From text to signal: evaluating LLMs for valid case identification in pharmacovigilance

Ewa Borowiack, Ewelina Sadowska, Joanna Konieczna, Monika Opałek, Iwona Kmicikiewicz, Damian Stachura, Artur Nowak
Evidence Prime, Krakow, Poland

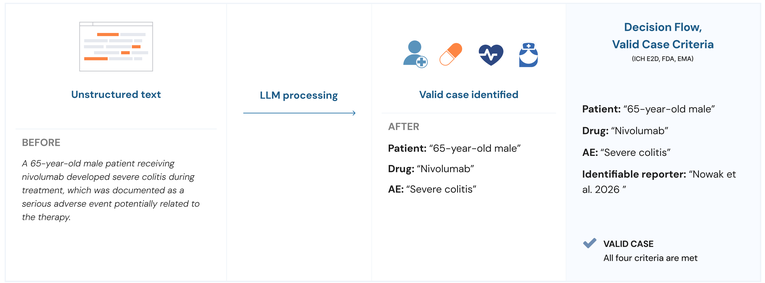
## Background

Pharmacovigillance (PV) requires rapid identification of valid individual case safety reports (ICSRs) within an expanding body of scientific literature. A valid case is defined by the presence of four minimum criteria: an identifiable patient, an identifiable reporter, a suspected medicinal product, and an adverse event, as established by regulatory guidelines (ICH E2D; FDA, EMA) [1,2]. Manual screening of literature to identify such cases is labor-intensive, time-consuming, and subject to inter-reviewer variability, particularly in complex reports such as case series or publications with implicit causality statements.

Recent advances in natural language processing, particularly large language models (LLMs), have demonstrated strong capabilities in information extraction from unstructured biomedical text [3]. In pharmacovigilance, early applications suggest that AI-assisted approaches can support case detection, coding, and triage, but challenges remain in ensuring traceability, consistency, and regulatory compliance

Specialized LLM agents may streamline literature screening by extracting case elements as structured, database-ready records with an auditable trail, including supporting quotes, document location, and concise rationale. Such structured outputs enable downstream standardization processes, including mapping to controlled vocabularies such as MedDRA [Fig 1].

**Transparent AI-assisted decision workflow** Fig 1




## Objective

In this study, we evaluated an LLM-agent pipeline for suspected drug identification as the first step toward automated extraction of all valid case elements, supporting scalable and transparent valid case identification in PV workflows. This study is performed as a part of the EU-funded LASER LLM project (FENG.01.01-IP.02-4479/23) that aims to extend Laser AI's automation capabilities by developing contextual, multilingual models capable of tagging all elements of a valid case and mapping them to standardized terminologies such as MedDRA as well as creating valid case identifier in the Laser AI.

## Methods

### Dataset development

A representative dataset of pharmacovigilance-relevant publications (n=71) was constructed to support identification of valid individual case safety reports (ICSRs). Eligible studies included case reports, case series, and selected randomized controlled trials.

Although the aim of the study was to assessed models for suspected drug identification, we decided to collect all data related to valid case identification necessary for future research. To ensure relevance for valid case identification, publications were required to report a suspected medical product (active substance or medical device, with or without brand name). No restrictions were applied to clinical domain to enhance diversity and generalizability.

**Data sources**

Records were collected from:

- expert-provided publications (including difficult cases),
- a focused PubMed query targeting adverse events in titles (e.g., adverse, toxicity, drug-induced, complication, overdose),
  with filters: free full text, case reports, English, humans, 2020–2025.

**Study selection**

Article selection was conducted in Laser AI using a two-stage screening process:

- title/abstract (TIAB) screening,
- full-text review.

Three reviewers participated in an iterative, batch-based process (≤100 records per batch) until the target dataset size (n=50?) was reached.

**Annotation process**

A structured extraction form was implemented in Laser AI based on predefined templates. Reviewers were trained using detailed guidelines and a pilot on 5 studies. Next, each record was annotated by a single reviewer (total n=3 reviewers). The extraction process leveraged:

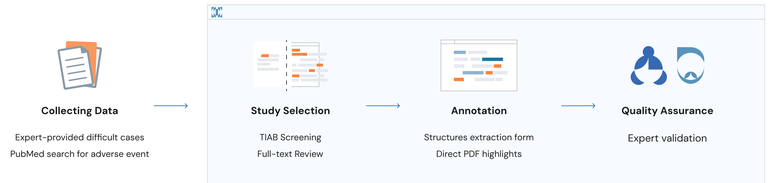
- direct highlighting of source text in PDFs,
- structured fields capturing:
  - extracted value,
  - reported value,
  - reviewer comment (used for reasoning).

If there was more than one valid case in a single record, each was treated as a distinct object, as shown in [Fig 3].

**Quality assurance**

All extracted data underwent domain expert validation to ensure consistency and accuracy of annotations.

**Database development process** Fig 2




Actions performed within the Laser AI platform

### Model development and evaluation

The training subset of the dataset was used to iteratively develop prompts and configure domain-specific LLM agents. Agents were designed to operate as a structured extraction pipeline, generating one standardized record per publication.

Model outputs were compared against human-annotated reference data to assess extraction accuracy. Quantitative evaluation was performed using F1 scores calculated on structured outputs. F1 score is defined as the harmonic mean of precision and recall, summarizing how well a model balances missed detections and false alar ms.

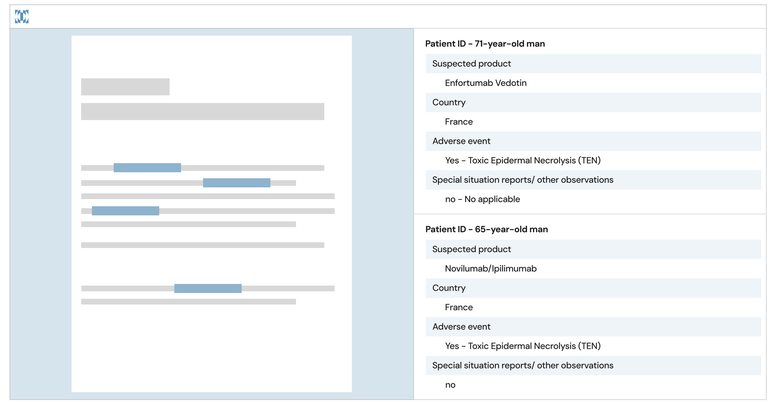
In addition, qualitative error analysis was conducted to systematically categorize discrepancies, including missing entities, incorrect aggregation of drug information, and conservative omission of implicitly reported suspected drugs.

## Results

Agents achieved an F1 score of 80% on the held-out test set (n=50), demonstrating reliable performance in identifying suspected drugs from pharmacovigilance literature.

Error analysis revealed that the majority of failures (82%) occurred in multi-patient case series, where not all patient–drug pairs were correctly captured. Additional failure modes included incorrect handling of combination therapies, with agents returning merged entities (e.g., "ipilimumab/nivolumab") instead of individual drugs, and conservative causality assessment, leading to omission of suspected drugs when attribution was implicit or uncertain. Also, model tagged to many interventions as suspected for adverse event.

**Data Annotation Dashboard Visualization** Fig 3




## Conclusion and future directions

Specialized LLM agents can reliably identify and tag suspected drugs in pharmacovigilance literature, achieving strong performance while producing structured, auditable outputs that support transparent and scalable case triage. By linking each extracted element to supporting evidence and rationale, the approach aligns with regulatory expectations for traceability and facilitates human verification.

Importantly, the results highlight both the potential and current limitations of AI-based extraction, particularly in complex scenarios such as multi-patient reports and implicit causality. These findings provide a clear direction for further refinement of extraction strategies and prompt design.

Future work will extend the pipeline to cover all remaining valid case elements, including patient, reporter, and adverse event, enabling full automation of valid case identification. Additional efforts will focus on improving handling of complex clinical narratives, expanding coverage to non-English publications, and incorporating multimodal inputs (e.g., tables, figures).

Integration with standardized medical ontologies (e.g., MedDRA) will further support downstream structuring and interoperability, ultimately enabling end-to-end, AI-assisted pharmacovigilance workflows for large-scale literature monitoring.

## References

[1] U.S. Food and Drug Administration. Postmarketing Adverse Event Reporting Compliance Program [Internet]. Silver Spring (MD): FDA; 2025 [cited 2026 Apr 7]. Available from: FDA website.
[2] European Medicines Agency, Heads of Medicines Agencies. Guideline on good pharmacovigilance practices (GVP): Annex I – Definitions (Rev 5) [Internet]. Amsterdam: EMA; 2024 Jul 26 [cited 2026 Apr 7]. Available from: EMA website.
[3] Walker VR, Schmitt CP, Wolfe MS, Nowak AJ, Kulesza K, Williams AR, Shin R, Cohen J, Burch D, Stout MD, Shipkowski KA, Rooney AA. Evaluation of a semi-automated data extraction tool for public health literature-based reviews: Dextr. Environ Int. 2022 Jan 15;159:107025. doi: 10.1016/j.envint.2021.107025. Epub 2021 Dec 14. PMID: 34920276.