PVLens: Enhancing Pharmacovigilance Through Automated Label Extraction

Jeffery L. Painter, MS, JD[1], Gregory E. Powell, PharmD, MBA[1]
Andrew Bate, PhD[2,3]
[1]GSK, Durham, NC, USA; [2]GSK, London, UK; [3]London School of Hygiene and Tropical Medicine, London, UK

Abstract

Reliable drug safety reference databases are essential for pharmacovigilance, yet existing resources like SIDER are outdated and static. We introduce PVLens, an automated system that extracts labeled safety information from FDA Structured Product Labels (SPLs) and maps terms to MedDRA. PVLens integrates automation with expert oversight through a web-based review tool. In validation against 97 drug labels, PVLens achieved an F1 score of 0.882, with high recall (0.983) and moderate precision (0.799). By offering a scalable, more accurate and continuously updated alternative to SIDER, PVLens enhances real-time pharamcovigilance with improved accuracy and contemporaneous insights.

Introduction

A clear understanding of known adverse effects, along with continuous surveillance for emerging safety concerns, is essential for patients, healthcare professionals, and pharmacovigilance (PV) scientists. Structured reference sets, such as FDA-approved drug labels, are critical for safety evaluation. One of the most widely used reference sets in PV is the FDA-approved drug label, which provides official safety and efficacy information for medicinal products in the US.

Despite their importance, no continuously updated, gold-standard resource exists for systematically accessing labeled drug events [1]. The Side Effect Resource (SIDER) has been widely used in drug discovery and PV workflows [2], yet it has not been updated since 2015, making it increasingly misaligned with current safety knowledge. Updates in MedDRA terminology have led to obsolete, incorrect, and misleading mappings, affecting drug safety assessments — for example, SIDER lists “urine output” as an adverse event, yet it was reclassified as a quantitative concept in the UMLS, falling outside SIDER's own original inclusion criteria. At least 40 similar terms have been reclassified. Additionally, SIDER’s reliance on PubChem IDs has introduced significant annotation errors, such as misclassifying Lescol with CID-124838623—an identifier now assigned to Remdesivir. These limitations undermine its reliability, particularly for newer drugs requiring rigorous safety monitoring.

Recent efforts have explored machine learning (ML)-based extraction methods, from BERT-based models to LLM-driven systems like OnSIDES and AskFDALabel [3, 4]. While scalable, these approaches currently struggle with MedDRA-specific mappings and hierarchical relationships [5]. For example, OnSIDES mapped 4,423 distinct MedDRA terms to adverse events and indications, whereas our pipeline mapped over 8,640 distinct terms, demonstrating a limitation in LLM abilities to robustly capture adverse event terminology. More broadly, these ML-based and term-matching systems attempt to infer product mappings without leveraging structured biomedical resources, such as UMLS RxNorm and MTHSPL (Metathesaurus SPL entries), introducing redundant complexity into the extraction process.

Rather than relying on ML-based substance identification, PVLens directly integrates UMLS and RxNorm mappings, eliminating the need for inferred mappings. Since RxNorm is derived from FDA Structured Product Label (SPL) data, it provides the most reliable linkage between drug substances and labeled adverse events. This structured, verifiable approach enhances accuracy, reduces ambiguity, and improves scalability for PV analytics.

To address these challenges, we introduce PVLens, an automated system that extracts safety data from FDA SPLs and maps terms to MedDRA, RxNorm, and SNOMED CT. PVLens processes SPL XML data using dictionary-based NLP and UMLS resources [6], while integrating Safety-Related Labeling Changes (SrLC) [7]. Optimized for large-scale processing, PVLens can extract and process over 50,000 SPLs in under an hour, ensuring frequent updates aligned with evolving safety information. PVLens is designed for periodic automated reprocessing of all SPLs, enabling

monthly or quarterly updates as new label data or terminology changes are released. The system checks for MedDRA updates using UMLS release notes and remaps extracted terms accordingly. This ensures that PVLens remains aligned with evolving regulatory terminology. While some commercial pharmacovigilance tools exist, they are proprietary and lack transparency, with evaluations revealing incomplete substance mappings. In contrast, PVLens will be released as an open-source resource, ensuring accessibility and reproducibility for PV analytics.

To assess its effectiveness, we conducted a structured validation study comparing PVLens' automated extractions to expert-reviewed annotations across 97 drug labels (n=63 post-2014 with documented labeled event changes, n=34 pre-2014)[1]. This study evaluates how well PVLens captures labeled safety information and aligns with expert judgment. Unlike SIDER, which has been widely cited but never formally validated, we believe PVLens provides the first benchmarked system for structured label extraction. Additionally, as no officially reviewed MedDRA mappings exist for post-2014 products, this evaluation represents the first structured comparison of extracted label content using a reproducible methodology.

Methods

System Overview and Data Extraction

PVLens employs a multi-phase extraction pipeline to ensure precise term identification and MedDRA mapping, as outlined in Figure 1. The process begins with a custom XML parser extracting the labeled sections of interest (i.e. indications, adverse events, black box warning) from SPLs (Step 1). Each label is then mapped to the UMLS using its Global Unique Identifier (GUID) by way of MTHSPL (Step 2), followed by RxNorm and SNOMED mappings to NDC codes (Step 3). Extracted text is mapped to MedDRA terms using NLP-based pattern matching (Step 4), and reported safety event dates are updated using SrLC (Step 5). The final outputs are merged into the PVLens repository, ensuring continuous updates for pharmacovigilance (Step 6).

The entire pipeline processes all prescription SPLs in under an hour, covering 5,358 distinct substances. To reduce noise in MedDRA mapping, generic phrases (e.g., adverse reaction) are filtered using a predefined stop word list[2], refined based on prior work [8].

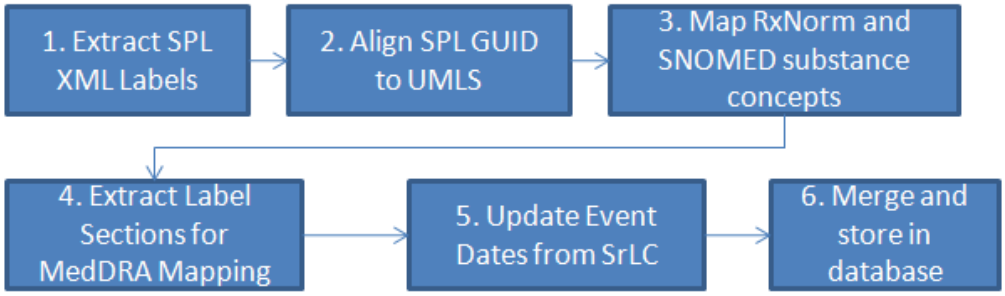

Figure 1: PVLens Processing Pipeline Overview. [GUID = Global Unique Identifier, SrLC = Safety-Related Label Change]

Study Design and Product Selection

To evaluate PVLens, we conducted a structured validation using independent review and adjudication of extracted terms. 97 distinct substance SPL labels were selected for review. Of these, 63 were approved post-2014 and had a labeled event change or black box warning added within two years post-approval, ensuring relevance for evaluating PVLens' ability to assist in the detection of emerging safety signals. The remaining 34 labels were selected at random and contained at least one mapped indication and one adverse event.

---

[1] Of the 250 labels initially selected for review, 97 have been completed and reported here. The review is ongoing, and the full data set will be made available once this process is finished.

[2] https://gist.github.com/sebleier/554280

Review and Adjudication Process

A total of 12 reviewers, including second- to fourth-year pharmacy students and a senior pharmacovigilance scientist, participated in the study. Each label was independently reviewed by two randomly assigned reviewers, with discrepancies on MedDRA term inclusion resolved by an expert adjudicator (a PharmD, PV scientist with 20+ years of experience). The 97-label review and adjudication process were completed within three weeks, requiring approximately 57 hours (excluding adjudication). The mean review time per label was 17.2 minutes, with a median of 28 minutes. Some complex labels required over 100 minutes for review due to their detailed safety content.

Beyond discrepancy resolution, reviewers could propose additional terms absent from the NLP-extracted outputs. These user-added terms were consolidated and reassessed by the adjudicator for final validation. Additionally, borderline or ambiguous mappings were flagged during review. The adjudicator applied a set of heuristics, prioritizing terms with clear MedDRA PT mappings, exclusion of vague descriptors (e.g. "abnormal test result"), and clinical relevance. If a term could not be confidently matched, it was excluded unless consensus could be reached through expert judgement.

Results

Reviewers were assigned labels based on availability, meaning that the study did not follow a paired review design. Consequently, traditional inter-rater reliability (IRR) metrics, such as Cohen's kappa, were not applicable since reviewers did not consistently evaluate the same set of labels. Instead, overall agreement across the 97 labels was measured, with independent reviewers agreeing on MedDRA term inclusion or exclusion 77% of the time. Adjudicator-reviewer agreement was consistently high, with a median of 91.3%, reinforcing the reliability of the review process. The limited number of false negatives (FN = 79, 1.3% of total terms reviewed) further supports the internal consistency of the review process and indicates that the vast majority of MedDRA term assignments were confirmed across multiple reviewers[3].

The algorithm was most effective in detecting adverse events, followed by indications, with black box warnings presenting the most challenges (Table 1). High recall across all categories suggests minimal missed MedDRA terms, though precision varied, indicating a tendency to over-include terms that required expert adjudication.

Table 1: Validation study results.

Category	TP	FP	FN	**Precision**	**Recall**	**F1-score**
Adverse Event	4,223	887	66	82.6%	98.5%	89.9%
Indication	170	110	10	60.7%	94.4%	73.9%
Black Box Warning	187	155	3	54.7%	98.4%	70.3%
Overall	4,580	1,152	79	**79.9%**	**98.3%**	**88.2%**

False Negative Analysis

To assess false negatives, reviewers could manually select or add terms to each section under review. During our final analysis, these user added terms were mapped to MedDRA using the same NLP approach in the overall pipeline. A term was then classified as a false negative only if it met three criteria: (1) it belonged to a valid semantic type[4], (2) it was not already mapped to a MedDRA synonym, and (3) it was not excluded as a stop word by PVLens.

[3] All data is available for inspection and review at https://github.com/jlpainter/AMIA2025/tree/main/pvlens/

[4] Initially, our goal was to try to reproduce the SIDER database and stay true to what has been made public about how the original authors constructed SIDER which lacks open-source code or documentation other than the original published manuscript. From the SIDER manuscript's supplementary materials, it contained a listing of the valid semantic types (Supplementary Table 1) they had selected for inclusion of adverse event terms, and we fully

Of 1,012 user-added terms, only 79 (7.8%) were successfully mapped to MedDRA Preferred Terms (PT) or Lower-Level Terms (LLT), comprising 3 black box warnings, 10 indications, and 66 adverse events. Most excluded false negatives were due to synonym recognition or classification differences, with 12 terms already captured by MedDRA synonyms and 12 classified as outside valid semantic types. Additionally, 9 terms were excluded based on predefined stopword filters, ensuring only clinically relevant terms were retained. The low false-negative rate suggests PVLens effectively captures relevant terms, with omissions primarily due to synonym variations or nuanced phrasing.

Discussion

PVLens serves as a continuously updated, transparent alternative to static resources like SIDER, enabling real-time tracking of evolving drug safety label information. By incorporating temporal tracking and the SrLC, PVLens enhances drug safety relevance. Future work will focus on improving precision, potentially by integrating LLMs for automated adjudication and refining term-matching approaches.

PVLens prioritizes high recall to ensure comprehensive capture of labeled adverse events (AEs), as missing critical AEs poses a greater risk than including extraneous terms. While recall exceeded 98%, precision was 80%, indicating that the algorithm captures more MedDRA PTs than required. Reviewer input confirmed only 79 terms were missed (3 black box warnings, 10 indications, 66 AEs), reinforcing strong MedDRA coverage. The lack of mandated vocabulary such as MedDRA in the SPL's adverse event sections introduces challenges in maintaining consistent term mappings. Standardizing the use of MedDRA at the point of label creation might enhance downstream analytics and signal detection, and PVLens highlights the value of enforcing such alignment for public safety.

For routine PV, these findings suggest that PVLens is a valuable resource requiring minimal further refinement. Its high recall ensures comprehensive capture of labeled AEs and indications, reducing the risk of missing critical safety information. While expert review remains necessary, PVLens provides a structured, scalable repository for US-approved products, supporting both routine screening and structured case evaluation.

PVLens minimizes false negatives (FNs), reducing the likelihood of missing critical AEs. While false positives (FPs) occur, they are manageable through rapid adjudication, a far more efficient process than manually reviewing every SPL. This significantly improves efficiency over traditional manual label reviews, which are time-consuming and inconsistently applied. Our performance metrics align with MITRE and FDA assessments of NLP techniques for AE extraction, which reported up to 79% F1 for MedDRA coding, compared to PVLens' 88.2% [9]. Similarly, NLP-based MedDRA annotation of drug labels have shown F1 scores ranging from 67%-79%, highlighting the inherent challenges of optimizing extraction performance [10].

Other initiatives, such as RS-ADR [11], integrate EHR and real-world data look to validate adverse drug reaction (ADR) signals, enhancing post-market surveillance. However, such approaches rely on retrospective clinical data, which is subject to reporting bias and coding inconsistencies. PVLens complements these efforts by offering a structured, up-to-date resource, ensuring a comprehensive repository of AEs and indications. Integrating structured label extractions with real-world validation can further strengthen PV.

Advancements in biomedical informatics offer new opportunities to refine PVLens beyond rule-based extractions. Future iterations will explore context-aware embeddings to improve term-matching and reduce FPs. Additionally, integrating semantic reasoning with UMLS, MedDRA, and RxNorm could enhance concept linkage across regulatory datasets.

---

replicated the same set of semantic types in this study. MedDRA terms can be linked to the semantic type definitions via the UMLS CUI (concept unique identifier) and the STN (Semantic Type ID) found within the MRSTY table in a UMLS extraction. Only MedDRA terms falling under the primary STN categories as used in SIDER are then included.

Harmonizing safety information across global regulatory bodies remains a challenge. PVLens' ontology-driven approach enables expansion beyond FDA SPLs to EMA, PMDA, and WHO reports, fostering global PV alignment. Future work includes adapting the extraction pipeline to accommodate different label structures and multilingual content, including Japanese and European regulatory formats. This expansion may require additional ontology mapping and label structure harmonization but is well-aligned with PVLens' modular architecture. To our knowledge, PVLens is the first validated comprehensive ADR database, systematically evaluated against human reviewers. By combining structured label extraction with AI-driven surveillance, PVLens lays the foundation for hybrid PV systems, integrating regulatory knowledge with real-world evidence. As agencies adopt AI-driven monitoring, PVLens offers a scalable, interpretable, and auditable framework for drug safety analytics.

Beyond pharmacovigilance, PVLens offers clinical utility by enabling integration into decision support systems. For example, its output can be used to inform drug safety alerts within EHRs, generate context-aware warnings during prescribing, or support structured patient safety reviews. Its transparent mappings also enhance reproducibility in clinical research and facilitate post-market label surveillance.

PVLens has an open database structure that can be easily integrated into existing health IT systems via standardized export formats (e.g., JSON, CSV) or through RESTful APIs. The modular architecture could also support future FHIR compatibility for embedding within clinical workflows or safety review dashboards.

Conclusion

Modern pharmacovigilance requires accurate, continuously updated, and validated resources. PVLens meets this need by extracting and mapping labeled safety events from FDA SPLs to MedDRA, RxNorm, and SNOMED. With high recall and solid precision, achieved through dictionary-based tokenization and context-aware filtering, PVLens enhances pharmacovigilance workflows and supports structured safety assessments.

Validation against human reviewers confirms PVLens' ability to comprehensively capture labeled safety information, making it well-suited for routine screening and case evaluation. While improving precision remains a priority, high recall minimizes the risk of missing critical safety events. Future work will focus on expanding beyond U.S. products and integrating LLMs and AI-driven techniques to refine term matching.

To our knowledge, PVLens is the first validated ADR database, systematically evaluated against human reviewers. By offering a continuously updated, reproducible repository of labeled indications and AEs, PVLens provides a scalable, evolving alternative. It strengthens the foundation for drug safety analytics, enhancing risk assessment and signal evaluation in real time.

Acknowledgements

We acknowledge the UNC Eshelman School of Pharmacy PharmD students for reviewing SPL data under supervision of Greg Powell, PharmD: Miranda Barker, Hibbah Ashraf, Megan Earnhart, Margaux Meilhac, Breannah Keys, Ravi Parekh, Emily Wu, Vicky Mei, Derek Bassett, Joshua Airgood, Ryan Le and Rowena Dzorvakpor.

Declarations

GSK covered all costs associated with the conduct of the study and the development of the manuscript and the decision to publish the manuscript. J.P., G.P, and A.B. are employed by GSK and hold financial equities. This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Smith JC, Denny JC, Chen Q, Nian H, Spickard Iii A, Rosenbloom ST, Miller RA. Lessons learned from developing a drug evidence base to support pharmacovigilance. Applied clinical informatics. 2013;4(04):596-617.

2. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic acids research. 2016 Jan 4;44(D1):D1075-9.

3. Tanaka Y, Chen HY, Belloni P, Gisladottir U, Kefeli J, Patterson J, Srinivasan A, Zietz M, Sirdeshmukh G, Berkowitz J, Brown KL. OnSIDES (ON-label SIDE effectS resource) database: extracting Adverse Drug Events from Drug Labels using Natural Language Processing Models. medRxiv. 2024 Mar 24:2024-03.

4. Wu L, Fang H, Qu Y, Xu J, Tong W. Leveraging FDA Labeling Documents and Large Language Model to Enhance Annotation, Profiling, and Classification of Drug Adverse Events with AskFDALabel. Drug Safety. 2025 Feb 20:1-1.

5. Dong G, Bate A, Haguinet F, Westman G, Dürlich L, Hviid A, Sessa M. Optimizing signal management in a vaccine adverse event reporting system: a proof-of-concept with COVID-19 vaccines using signs, symptoms, and natural language processing. Drug Safety. 2024 Feb;47(2):173-82.

6. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004 Jan 1;32(suppl_1):D267-70.

7. Munesh G, Bhavaraju ML. Regulatory framework for USFDA regulated drug product labeling update. Pharmaceutical Sciences Asia. 2024 Jan 1;51(1)

8. Demner-Fushman D, Shooshan SE, Rodriguez L, Aronson AR, Lang F, Rogers W, Roberts K, Tonning J. A dataset of 200 structured product labels annotated for adverse drug reactions. Scientific data. 2018 Jan 30;5(1):1-8.

9. Bayer S, Clark C, Dang O, Aberdeen J, Brajovic S, Swank K, Hirschman L, Ball R. ADE eval: an evaluation of text processing systems for adverse event extraction from drug labels for pharmacovigilance. Drug safety. 2021 Jan;44:83-94.

10. Ly T, Pamer C, Dang O, Brajovic S, Haider S, Botsis T, Milward D, Winter A, Lu S, Ball R. Evaluation of Natural Language Processing (NLP) systems to annotate drug product labeling with MedDRA terminology. Journal of biomedical informatics. 2018 Jul 1;83:73-86.

11. Lee S, Lee JH, Kim GJ, Kim JY, Shin H, Ko I, Choe S, Kim JH. A data-driven reference standard for adverse drug reaction (RS-ADR) signal assessment: development and validation. Journal of Medical Internet Research. 2022 Oct 6;24(10):e35464.