Clinical Evidence and Patient Reviews: A Linked Dataset for Antidepressant Side Effects

Steven Au
Independent Researcher
steventinwing@gmail.com

Abstract

Clinical sources and patient-authored reviews often describe antidepressant side effects in different ways, but these differences are rarely measured directly. We present ClinPeer-AE, a linked dataset for comparing side-effect evidence from PubMed, ClinicalTrials.gov, WebMD, and Drugs.com while preserving source identity. Across five widely prescribed antidepressants, we find low overlap between clinical and peer sources, large differences in relative emphasis, and evidence that many peer-only effects also appear in U.S. Food and Drug Administration Adverse Event Reporting System (FAERS) reports. These findings suggest that patient reviews provide useful context about recurring medication experiences and offer a complementary view of how side effects are described outside formal clinical settings.

1 Introduction

Clinical trials and patient review platforms describe antidepressant treatment in different ways. Trials use controlled designs and standardized adverse-event tables. Reviews show how people talk about treatment in everyday settings, including side effects, discontinuation, and comparisons with other drugs. Both sources matter, but they do not always use the same language.

This paper asks whether those differences can be measured without flattening the sources together. If reviews use different terms, mention effects at different rates, or connect them to specific treatment contexts, that information should remain visible.

We address this with **ClinPeer-AE**[1], a linked dataset that keeps clinical and peer evidence separate while making them comparable. The contribution is not a new adverse-event extractor. Instead, we show how much depends on tracking source, context, and role, especially in reviews where the same phrase may describe an indication, a benefit, a withdrawal symptom, or an on-treatment adverse effect.

Our contributions are:

1. A linked dataset for comparing clinical and patient-authored side-effect evidence for five antidepressants.
2. A set of source-level analyses covering effect coverage, mention salience, condition context, and evidence support.
3. An extraction-quality and sensitivity analysis showing why role classification matters for patient reviews.

2 Related Work

Research on adverse drug effects has been supported by a strong ecosystem of structured resources, benchmark datasets, and biomedical extraction tools. The Side Effect Resource (SIDER) provides curated drug–side-effect associations derived from product labels (Kuhn et al., 2016), while the Medical Dictionary for Regulatory Activities (MedDRA) offers standardized terminology for adverse event reporting in regulatory contexts (Brown et al., 1999). Benchmark corpora such as the Adverse Drug Event (ADE) corpus from medical case reports (Gurulingappa et al., 2012), the CSIRO Adverse Drug Event Corpus (CADEC) from patient forum posts (Karimi et al., 2015), and social media datasets developed for pharmacovigilance (Sarker and Gonzalez, 2017; Feyisetan et al., 2020) have helped define the task in terms of entity recognition, relation extraction, and concept normalization. More broadly, shared tasks in biomedical event extraction have helped shape information extraction methods for clinical and biomedical text (Pyysalo et al., 2012; Kim et al., 2011), and tools such as SciSpaCy have made it easier to build practical pipelines for biomedical parsing and named entity recognition (Neumann et al., 2019).

[1] https://github.com/Prowo/ClinPeer-AE

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 656–664

Adverse-effect evidence, however, is distributed across sources that differ substantially in purpose, language, and reporting style. Clinical trials, regulatory labels, case reports, and patient-authored reviews do not simply present the same information in different formats. They foreground different aspects of medication experience and often vary in granularity, framing, and salience. Clinical trial adverse event tables summarize outcomes observed under controlled conditions, whereas patients often describe effects in more subjective and functionally meaningful terms, such as emotional numbing, cognitive fog, or sexual dysfunction. Prior work has shown meaningful differences between these perspectives, especially for psychiatric medications. The Psychiatric Treatment Adverse Reactions dataset (PsyTAR) provides one of the first dedicated annotated datasets for psychiatric treatment reviews (Zolnoori et al., 2019), and pharmacovigilance studies have shown that social media can surface safety signals that may be underrepresented in formal reporting systems. However, most prior work treats patient-authored text as a supplementary signal for adverse event detection rather than as a source with distinct descriptive and evidential properties.

Recent advances in pretrained and instruction-tuned language models have also improved zero-shot and weakly supervised analysis of medical text. Bidirectional Encoder Representations from Transformers (BERT) family models, including DeBERTa-v3 (Liu et al., 2023) and ModernBERT (Warner et al., 2025), along with instruction-tuned models such as Gemma 4 (Google DeepMind, 2026), make it increasingly feasible to process patient reviews without large task-specific training sets. This is particularly important in patient-authored narratives, where systems must distinguish true adverse effects from related but different phenomena, including treatment indications, perceived benefits, withdrawal symptoms, prior medication history, and background conditions. In this setting, role classification is often necessary to determine how a mention should be interpreted.

3 Data

We assemble a multi-source corpus covering five widely prescribed antidepressants: bupropion, escitalopram, fluoxetine, sertraline, and venlafaxine. The corpus is organized into two source families (Table 1). The *clinical family* comprises PubMed (abstracts and study descriptions) and ClinicalTrials.gov (structured adverse-event tables when available). The *peer family* comprises WebMD and Drugs.com, both of which host patient-authored narratives with condition and rating metadata. All downstream analyses preserve this family label rather than pooling sources before comparison.

Table 1: Corpus composition for the five focal antidepressants.

Source	Docs	Claims	Drugs
PubMed	346	1,871	5
ClinicalTrials.gov	51	29	3
WebMD	18,510	34,670	5
Drugs.com	7,693	39,812	5
Total	**26,600**	**76,382**	–

We use "clinical" to mean formal biomedical evidence broadly, not clinical trials specifically, since PubMed captures a wider range of study designs than trials alone. The released resource is accordingly named **ClinPeer-AE** to reflect this scope. We restrict the corpus to structured review platforms (WebMD, Drugs.com) and biomedical databases (PubMed, ClinicalTrials.gov), excluding social-media sources such as Reddit. Across these four sources, the corpus contains 26,600 documents and 76,382 source-level drug–side-effect claims; downstream metrics aggregate by drug, effect, and source family.

4 Method

4.1 Claim Extraction

Documents are ingested into a shared DuckDB store. A *claim* is an extracted tuple

$$c = (d, e, s, f, r, a, q, x),$$

where d is the drug, e is the effect, s is the source, f is the source family, r is the contextual role, a is the attribution tier, q is confidence, and x is the evidence span. The downstream graph uses the normalized drug–effect pair as the edge key, but retains the source family, source identifier, role, attribution tier, count, and confidence annotations.

The extractor is hybrid rather than a single SciSpaCy model. SciSpaCy provides sentence segmentation, tokenization, biomedical named-entity candidates, and dependency parses. Project-specific

lexicons provide antidepressant aliases and side-effect aliases; these are curated into canonical identifiers before graph construction. Candidate pairs are generated differently by source type. ClinicalTrials.gov adverse-event tables are imported as structured drug–effect rows with available frequency metadata. PubMed abstracts and patient reviews are segmented into sentences; a drug alias and an effect candidate must occur in the same sentence or in a short adjacent-window context, unless the source metadata already fixes the reviewed drug. Dependency distance, trigger terms, negation cues, and lexical match quality determine the confidence score. Each candidate is then assigned an attribution tier encoding how strongly a source frames the effect as medication-related (direct causation > temporal association > co-occurrence > unclear).

To illustrate, consider the sentence "Sertraline caused severe insomnia after two weeks." SciSpaCy segments and tokenizes the input, tags "sertraline" as a drug candidate via the `en_core_sci_md` model, and identifies "insomnia" as a side-effect candidate. The project lexicon maps both to canonical identifiers. The dependency parse links the two entities through the causal trigger "caused," yielding a direct-causation claim for sertraline, insomnia, Drugs.com, peer source, adverse-effect role, confidence 0.92, and the evidence span "caused severe insomnia." By contrast, "Sertraline helped my panic attacks" receives role *benefit* and does not enter the adverse-effect claim set.

For peer sources, we add a review role-classification layer that assigns each candidate drug–effect pair to one of seven context roles: *indication*, *benefit*, *adverse effect*, *withdrawal effect*, *no effect*, *comparator/prior medication*, or *background condition*. Only `adverse_effect` claims pass the role gate in gated configurations. Withdrawal effects are retained as metadata rather than discarded, as discontinuation phenomena may be analytically useful but should not be conflated with on-drug AEs.

Table 2 shows schematic examples of how the role layer changes emitted claims. These examples are paraphrased patterns, not quoted patient reviews.

4.2 Graph Representation

ClinPeer-AE models drugs and side effects as graph nodes, preserving source-family branching at the edge level. Clinical and peer evidence are not merged; the internal graph stores source-specific counts and supports comparative views: clinical-only, peer-only, shared, and high-divergence effects. For each drug, edges to side-effect nodes are annotated with claim counts, average confidence, divergence scores, and rank inversions. An optional FAERS branch adds pharmacovigilance report counts as a third evidence family.

The combined graph for all five drugs contains 87 nodes (5 drug + 82 unique side effects) and 732 branched edges. This design makes source differences measurable without discarding source information.

4.3 Metrics and Hypotheses

We organize the evaluation around four hypothesis tests.

H1: Coverage distinctness. Peer and clinical AE sets have low overlap. We measure this with the Jaccard index (shared / union) and bootstrap 95% CIs. H1 is supported when the upper CI falls below 0.5.

H2: Saliency divergence. Clinical trial AE frequency does not predict peer mention rates. For each shared effect, we compute a *saliency ratio* = peer_mention_count / clinical_trial_frequency_pct. A ratio of 1.0 indicates parity; higher ratios indicate greater relative peer mention salience. We test whether median ratios significantly exceed 1.0 using a Wilcoxon signed-rank test.

H3: Condition-stratified AE profiles. The same drug has different AE mention distributions depending on the treated condition (Depression, Anxiety, OCD, Panic Disorder, PTSD). We test independence using chi-squared contingency tests.

H4: Evidence support by divergence level. Peer-only (high-divergence) effects should carry different evidence backing than shared (low-divergence) effects. We compute a composite support score for each effect combining attribution tier strength (40%), independent peer document support (35%), cross-source peer agreement (25%), reviewer metadata (+0.5), and FAERS corroboration (+0.5). Because most score components are computed over extracted claims, this score is a comparative triage proxy rather than an independent measure of clinical truth. We compare high-divergence ($\geq$0.7) vs. low-divergence ($<$0.3) effects using the Mann–Whitney U test, with Jensen–Shannon divergence and conflict density used as supplementary divergence metrics.

Table 2: Schematic claim extraction examples. The role layer prevents indications, benefits, withdrawal effects, and prior-medication experiences from being counted as ordinary on-treatment adverse effects.

Source pattern	Evidence pattern	Assigned role	Emitted AE claim
ClinicalTrials.gov table	Nausea reported in the sertraline arm with trial frequency	adverse event	(sertraline, nausea)
Patient review	Escitalopram caused insomnia after dose increase	adverse effect	(escitalopram, insomnia)
Patient review	Sertraline helped panic attacks	benefit	none
Patient review	Brain zaps after stopping venlafaxine	withdrawal effect	metadata only
Patient review	Bupropion replaced a prior drug that caused fatigue	comparator/prior medication	none

5 Results

5.1 Extraction Quality

We evaluate extraction on a locked test of 29 Drugs.com reviews and a larger PsyTAR review set, with detailed scores reported in Appendix A. The locked set was produced by one annotator with AI-assisted candidate drafting followed by manual adjudication. Because we did not compute inter-annotator agreement, these scores should be read as a development-quality reliability estimate rather than a final benchmark.

A substantial share of the low F1 comes from role confusion, not only missed effect terms. In the locked test, many false positives are benefit or withdrawal mentions that the extractor counted as adverse effects. This is a central limitation: the downstream comparisons should be read as aggregate analyses over noisy extracted claims, not as verified lists of individual side effects.

We also evaluate separate role-classification baselines. Table 3 shows cross-domain results on PsyTAR review windows, and Appendix B gives additional Gold 120 results. The best local language-model baseline performs much better on adverse-effect role classification than the default extractor, which suggests a clear path for improving the pipeline. It is not yet integrated into the main experiments.

As a stress test, we re-run the peer pipeline with the deterministic role gate enabled. This removes most peer claims, but H1, H2, and H3 remain supported for all five drugs; H4 is more sensitive because its score depends directly on claim volume. This does not solve the extraction problem, but it suggests the main coverage and salience patterns are not driven only by obvious non-adverse review mentions.

Table 3: Cross-domain role classification on 300 PsyTAR overlap windows.

Model	Type	P	R	F1	Acc_7
Gemma 4 E4B	Causal LM	**.893**	.856	**.874**	.573
TF-IDF + LR	Trained	.770	**.966**	.857	**.603**
ClinicalBERT + LR	Trained	.812	.712	.759	.503
BioBERT + LR	Trained	.891	.616	.729	.483
DeBERTa-v3 (NLI)	Zero-shot	.606	.822	.698	.490
TinyLlama 1.1B	Causal LM	.470	.760	.581	.373
ModernBERT (NLI)	Zero-shot	.604	.199	.299	.240
Gemma 4 E2B	Causal LM	.804	.870	.836	.563

Table 4: Coverage distinctness (H1). All drugs show low clinical–peer overlap.

Drug	Clin	Peer	Shared	Jaccard	CI_{95}
Sertraline	102	74	27	.181	[.121, .242]
Fluoxetine	200	68	31	.131	[.089, .177]
Escitalopram	162	72	28	.136	[.092, .184]
Venlafaxine	118	68	24	.148	[.099, .204]
Bupropion	144	67	25	.134	[.086, .188]

5.2 H1: Coverage Distinctness

Table 4 presents coverage results. H1 is supported for all five drugs at the extracted-claim level: Jaccard overlap ranges from 0.13 to 0.18, with upper 95% CIs never exceeding 0.24. At least 76% of the effect space is non-overlapping even accounting for sampling variation. Conversely, 54–65% of peer-reported effects are absent from clinical AE tables.

5.3 H2: Saliency Divergence

Table 5 reports saliency results. H2 is supported for all five drugs (Wilcoxon $p < 10^{-6}$). Patients discuss shared side effects 17–96$\times$ more than clinical trial frequency would predict (Figure 2). Spearman rank correlation is significant for only one drug (fluoxetine), suggesting that saliency ratios capture information not visible in simple rank agreement.

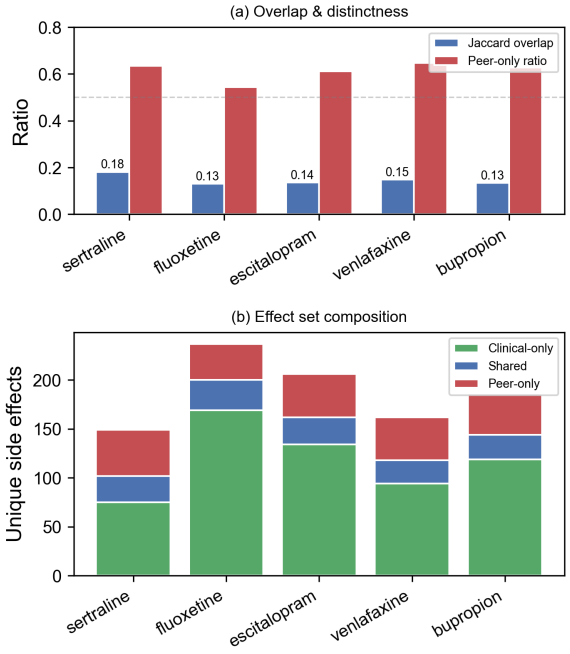

Figure 1: Overlap and effect-set composition by drug. Light regions denote peer-only effects; dark regions denote clinical-only effects.

Table 5: Saliency divergence (H2). Median saliency ratios and Wilcoxon signed-rank tests.

Drug	Shared	Med. ratio	p (Wilcoxon)
Sertraline	28	96.2×	3.7e-09
Fluoxetine	31	41.0×	1.4e-09
Escitalopram	29	56.7×	1.3e-06
Venlafaxine	25	16.6×	6.0e-08
Bupropion	26	68.8×	1.5e-08

5.4 H3: Condition-Stratified AE Profiles

Table 6 presents condition-stratified results. Four drugs show strong evidence of condition-stratified AE profiles; venlafaxine is nominally significant but treated as marginal ($p = 0.046$). Overall, the same drug can yield different AE mention profiles depending on the treated condition.

Notable findings include PTSD patients on venlafaxine reporting weight gain at 24% (vs. 2.9% for depression), bupropion-associated erectile dysfunction concentrated in PTSD reviews, and intrusive thoughts appearing primarily in OCD reviews for sertraline (10.3% vs. $\leq$0.7% for other conditions). Clinical trials typically aggregate AEs across indications; these results suggest that review-derived AE profiles can vary by treatment context. They should not be interpreted as causal estimates of condition-specific drug risk.

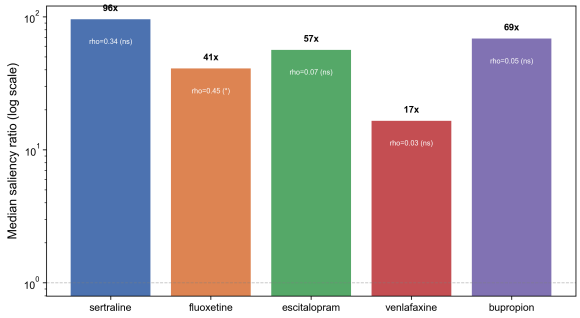

Figure 2: Median saliency ratio by drug (log scale). Dashed line at ratio = 1 (parity).

Table 6: Condition-stratified AE divergence (H3). Chi-squared tests of AE independence across treated conditions.

Drug	p	Top divergent effect
Sertraline	<0.001	intrusive thoughts (OCD 10.3% vs. Anxiety 0.7%)
Fluoxetine	<0.001	panic attacks (Anxiety 13.7% vs. Depression 3.4%)
Escitalopram	<0.001	panic attacks (PTSD 16.2% vs. Depression 4.5%)
Venlafaxine	0.046	weight gain (PTSD 24.0% vs. Depression 2.9%)
Bupropion	<0.001	erectile dysfunction (PTSD 13.5% vs. 0%)

5.5 H4: Evidence Support by Divergence Level

Table 7 presents H4 results. The hypothesis is supported for 4/5 drugs, but in the *opposite* direction: peer-only (high-divergence) effects receive lower support scores than shared (low-divergence) effects. The delta is consistently negative (-0.14 to -0.38). Because the score draws on extracted claim volume and pipeline-derived attribution, this result is best read as a suggestive comparative pattern rather than independent validation of evidential strength.

This finding is useful mainly for triage. It suggests three things: (1) divergence correlates with weaker measured support in this graph; (2) not all peer-only effects are equal, since some (e.g., *panic attack*, *brain zaps*) have higher support because they appear across platforms and in FAERS, while others may reflect extraction or reporting noise; (3) support scoring can prioritize which peer-only effects merit manual review and follow-up analysis.

FAERS triangulation. FAERS data (2021–2025, 36,826 cases) shows that 42–56% of peer-only effects also appear in suspect-drug pharmacovigilance reports, with the highest rate for venlafaxine (55.9%) and the lowest for fluoxetine (42.5%). This is external triangulation, not causal validation: FAERS is spontaneous-report data and can share reporting biases with review platforms. Still, the overlap suggests that many peer-only effects are

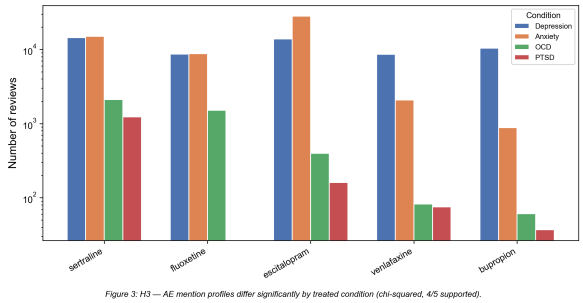

Figure 3: Condition-stratified AE review counts by drug.

Table 7: Evidence support by divergence level (H4). Peer-only effects receive consistently lower support scores than shared effects.

Drug	Peer-only	Shared	Δ	p
Sertraline	52	21	−0.27	0.0008
Fluoxetine	40	28	−0.38	4.9e-06
Escitalopram	60	12	−0.36	0.0003
Venlafaxine	59	9	−0.14	0.150
Bupropion	52	15	−0.22	0.011

not isolated review artifacts, while the support gap shows they remain thinner than shared effects on average.

Overall, most hypothesis tests reach nominal statistical significance across the five drugs. The main exception is venlafaxine under H4 ($p = 0.15$), where the direction is consistent but the sample of shared effects is small ($N = 9$). The venlafaxine H3 result is significant but marginal and should not be overinterpreted.

6 Discussion

The results suggest that differences between formal clinical evidence and patient reviews are not entirely random. Across five antidepressants, peer sources provide broader side-effect coverage and include recurring experiential terms such as brain zaps, emotional blunting, and intrusive thoughts that are weakly represented in clinical trial AE tables.

The H4 result runs counter to our initial expectation. We expected peer-only effects with high divergence to have stronger converging evidence, such as multiple independent reviews, cross-platform support, or FAERS corroboration. Instead, shared effects receive higher support scores. This pattern is suggestive but not definitive, since the score itself depends on extracted claim volume. It mainly shows that effects documented by both clinical and peer sources tend to accumulate more evidence in this graph. Peer-only effects often also appear in FAERS (42–56%), but they are thinner in this evidence base on average, making the score useful for triage rather than clinical adjudication.

Implications for extraction. The findings point to a practical need in patient-review analysis: extraction systems should model role and source explicitly. A simple drug–effect label cannot tell whether a phrase describes an indication, a benefit, a withdrawal symptom, a current side effect, or an experience with a prior medication. Evaluation should also keep clinical and review language separate enough to see where systems fail. ClinPeer-AE provides a testbed for that kind of role-aware extraction and source-level comparison.

One practical implication is that early source pooling hides the differences this paper is trying to measure. Keeping the sources separate makes those differences visible.

Evidence asymmetry. The peer corpus is substantially larger than the clinical-trial portion, and some observed divergence reflects study-design and reporting differences rather than purely biological difference. Clinical trials use strict inclusion criteria, controlled dosing, and structured AE capture; patient reviews reflect heterogeneous populations, variable adherence, and unstructured narrative. We discuss this as a limitation below, not as a flaw; the framework is designed to measure observable source differences, not to adjudicate biological truth.

7 Conclusion

ClinPeer-AE shows that clinical evidence and patient-authored reviews can be compared without merging them into one evidence pool. Across five antidepressants, the two source families differ in coverage, salience, condition context, and measured support. The main value of the dataset is therefore not a definitive list of side effects, but a reproducible way to study how adverse-event language and evidence vary across reporting settings.

Future work should improve extraction quality, add multi-annotator evaluation, expand clinical source coverage, and integrate stronger role classification before using peer-only effects as pharmacovigilance signals.

8 Limitations

Extraction F1 ranges from 0.11 to 0.21, primarily due to role confusion, including benefit and withdrawal mentions misclassified as AEs. The comparative framework does not eliminate this problem because source families have different language and therefore different error profiles. The Gemma 4 role classifier offers a clear improvement path, but is not yet integrated into the default pipeline.

The locked evaluation set was produced by a single annotator with AI-assisted candidate drafting and manual adjudication. No inter-annotator agreement was computed. The extraction scores therefore quantify performance against a useful but limited gold set rather than a mature shared-task benchmark.

We compare the production extractor to a lexicon baseline and a deterministic role gate, but we do not benchmark against strong supervised or generative extraction systems. A single extraction pipeline can also favor one source family over another, since clinical tables, PubMed abstracts, and patient reviews use different language. The NLP component should therefore be read as an enabling pipeline for source-aware analysis, not as a competitive extraction contribution.

WebMD and Drugs.com contain far more documents than ClinicalTrials.gov. Volume normalization helps but does not eliminate the asymmetry, and future work should incorporate additional clinical sources.

Patient reviews capture association, not causation, and FAERS reports reflect spontaneous reporting rather than controlled observation. The framework measures observable source differences; it does not establish which source is correct.

Patient reviews come from a self-selected population that may overrepresent negative experiences. The exclusion of Reddit also limits generalizability to broader peer discourse.

Five drugs across four hypotheses yield 20 tests without family-wise correction. Individual p-values should therefore be interpreted cautiously. Chi-squared tests for rare conditions such as OCD and PTSD also involve small cell sizes, and the venlafaxine H3 result ($p = 0.046$) should be treated as marginal.

9 Ethics Statement

This work analyzes publicly available evidence sources (PubMed, ClinicalTrials.gov, WebMD, Drugs.com, and FAERS) without identity linkage across platforms. Patient-review evidence is treated as observational and self-reported, not as causal ground truth. The goal is to quantify source-specific differences between formal evidence and patient-authored reviews in an auditable, reproducible manner, not to replace clinical evidence or to make treatment recommendations. All data is used in aggregate; no individual reviews are quoted or identifiable.

References

Elaine G. Brown, L. Wood, and S. Wood. 1999. The medical dictionary for regulatory activities (MedDRA). *Drug Safety*, 20(2):109–117.

Oluwaseun Feyisetan, Tom Diethe, and Suzanne Kasarda. 2020. Privacy-preserving word representations for improving pharmaceutical post-market surveillance. In *Proceedings of the 3rd Clinical NLP Workshop*, pages 79–85.

Google DeepMind. 2026. Gemma 4. Technical Report.

Harsha Gurulingappa, Anand M. Rajput, Alasdair Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support information extraction for adverse drug effects. In *Proceedings of BioNLP*, pages 73–80.

Sarvnaz Karimi, Alvaro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 1–6, Portland, Oregon, USA. Association for Computational Linguistics.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2016. The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(D1):D1075–D1079.

Pengfei Liu, Yuan Cao, Min Lin, Yijun Li, and Xudong Liang. 2023. DeBERTa-v3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. SciSpaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Overview of the ID, EPI and REL tasks of BioNLP shared task 2011. *BMC Bioinformatics*, 13(Suppl 11):S2.

Abeed Sarker and Graciela Gonzalez. 2017. A corpus for mining drug effects from social media. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 282–289.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Maryam Zolnoori, Kin Wah Fung, Timothy B. Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Nilay D. Shah, Yi Shuan Shirley Wu, Christina E. Eldredge, Jake Luo, Mike Conway, Jiaxi Zhu, Soo Kyung Park, Kelly Xu, and Hamideh Moayyed. 2019. The psytar dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. *Data in Brief*, 24:103838.

A Extraction Performance

This section gives the numeric extraction scores behind Section 5.1. The table is included as diagnostic context because the main analyses depend on extracted claims, but the detailed extractor comparison is not itself the central result.

Table 8: Extraction performance by configuration.

Config	Set	P	R	F1
scispacy_hybrid	Locked (29)	.174	.268	.211
scispacy_hybrid	PsyTAR (647)	.259	.071	.111
baseline lexicon	PsyTAR (647)	.256	.072	.112
scispacy + rules	PsyTAR (647)	.348	.005	.009

B Role Classification Baselines

Table 9 reports Gold 120 results for the seven-role classification task (locked_test + hard_case_dev splits). The PsyTAR cross-domain comparison is reported in the main paper in Table 3. "Trained" classifiers use PsyTAR overlap data; "Zero-shot" models require no task-specific training.

C Cross-Drug Peer-Only Effects

This section lists the recurring peer-only effects discussed briefly in the main results. It is kept in the appendix so the main paper can focus on the aggregate tests rather than a long effect inventory.

Nineteen effects are peer-only across all five drugs: *panic attack*, *brain zaps*, *emotional blunting*, *discontinuation syndrome*, *depersonalization*, *erectile dysfunction*, *excessive yawning*, *intrusive thoughts*, *anxiety attack*, *jaw clenching*, *numbness*, *back pain*, *muscle spasms*, *shaking*, *brain fog*, *delayed ejaculation*, *night sweats*, *confusion*, and *stomach pain*. An additional 13 effects appear in 4/5 drugs, including *tinnitus* and *agitation*. At the aggregate level, these terms recur across peer sources; however, individual effects still require manual and clinical validation before being treated as safety signals.

D Support Score Components

The composite support score is computed as:

- **Attribution tier** (40%): direct causation (1.0) $>$ temporal association (0.75) $>$ co-occurrence (0.50) $>$ unclear (0.25).
- **Independent peer documents** (35%): $\log(1+$ distinct_review_count$) \times 10$.
- **Cross-source support** (25%): binary flag for presence in both WebMD and Drugs.com.
- **Metadata support** (+0.5): review includes duration, condition, or demographics.
- **FAERS support** (+0.5): drug–effect pair appears in FAERS suspect-drug reports.

Total score range: 0 to 2.0.

E Supplemental: Attribution Language Comparison

The original H4 formulation compared attribution language between clinical and peer sources. All five drugs showed significant differences ($p < 10^{-36}$), but the effect sizes were small (Cliff's δ 0.07–0.20) and largely reflected genre differences: clinical trial AE tables list effects statistically ("treatment-emergent adverse event") while patient reviews use causal framing ("this drug caused nausea"). This result is technically significant but substantively uninformative and is reported here as supplemental.

Table 9: Review role classification on Gold 120 (Drugs.com reviews, 7 roles). Adverse-effect binary P/R/F1 and 7-class accuracy.

Model	**Type**	**Split**	**N**	**Adv P**	**Adv R**	**Adv F1**	**Acc$_7$**
Rules v1	Deterministic	pilot	148	1.00	.375	.545	.433
Gemma 4 E4B	Causal LM	locked_test	261	.889	.857	.873	.628
Gemma 4 E4B	Causal LM	hard_case_dev	367	.920	.896	.907	.676
Gemma 4 E2B	Causal LM	locked_test	261	.691	.839	.758	.628
Gemma 4 E2B	Causal LM	hard_case_dev	367	.803	.852	.827	.681
TF-IDF + LR	Trained	locked_test	261	.533	.429	.475	.448
TF-IDF + LR	Trained	hard_case_dev	367	.625	.652	.638	.425
ClinicalBERT + LR	Trained	locked_test	261	.326	.268	.294	.310
ClinicalBERT + LR	Trained	hard_case_dev	367	.637	.443	.523	.395
BioBERT + LR	Trained	locked_test	261	.325	.232	.271	.276
BioBERT + LR	Trained	hard_case_dev	367	.657	.565	.607	.373
DeBERTa-v3 (NLI)	Zero-shot	locked_test	261	.341	.821	.482	.391
DeBERTa-v3 (NLI)	Zero-shot	hard_case_dev	367	.512	.730	.602	.463
ModernBERT (NLI)	Zero-shot	locked_test	261	.377	.929	.536	.341
ModernBERT (NLI)	Zero-shot	hard_case_dev	367	.478	.748	.583	.436
TinyLlama 1.1B	Causal LM	locked_test	261	.189	.696	.298	.180
TinyLlama 1.1B	Causal LM	hard_case_dev	367	.308	.809	.446	.267