JURNAL PHARMACTIVE

JURNAL PHARMACTIVE | E-ISSN 2829-3444
Vol. 5 No 1 – April 2026 | https://s.id/jurnalpharmactive
DOI : https://doi.org/10.64036/pharmactive.v4i1
Publishing : LPPM Institut Teknologi dan Kesehatan Bintang Persada

NATURAL LANGUAGE PROCESSING UNTUK EKSTRAKSI INFORMASI ADVERSE DRUG REACTIONS DARI ELECTRONIC HEALTH RECORDS: SYSTEMATIC REVIEW

Riska Wiradarma[1], I Gede Irvan Pramanta Andika[2], Dody May Arfian[3]

[1]Program Studi Informatika, Fakultas Teknologi, Institut Teknologi dan Kesehatan Bintang Persada
Jalan Gatot Subroto, Denpasar, Indonesia
[2]Program Studi Informatika, Fakultas Teknologi, Institut Teknologi dan Kesehatan Bintang Persada
Jalan Gatot Subroto, Denpasar, Indonesia

e-mail: riska.wiradarma@gmail.com[1], irvanpramanta@gmail.com[2], dodymayarfian@gmail.com[3]

Received : January, 2026	Accepted : Maret, 2026	Published : April, 2026

Abstract

Adverse drug events (ADEs) contribute to 5-10% of hospitalizations and cost approximately USD 30 billion annually, yet spontaneous reporting systems capture only 5-10% of actual ADEs due to severe underreporting. This systematic review analyzed 60 peer-reviewed studies (2019-2025) on natural language processing (NLP) methods for extracting ADE information from electronic health record (EHR) clinical notes, following PRISMA 2020 guidelines across five databases (PubMed, IEEE Xplore, ACL Anthology, Scopus, Web of Science). Results demonstrate that transformer-based models, particularly BioBERT and ClinicalBERT, represent the state-of-the-art with F1-scores of 0.85-0.92 on benchmark datasets (n2c2 2018, MIMIC-III, MADE1.0), significantly outperforming rule-based systems (+15-20%) and traditional machine learning methods (+8-12%). Domain-specific pre-training on clinical text proved crucial, improving performance by 3-5% over general BERT models. However, critical challenges persist: negation and speculation detection (30-40% of medical mentions require contextual disambiguation), temporal reasoning for determining ADE onset relative to drug exposure, ambiguous medical abbreviation resolution, and causality assessment. A significant lab-to-clinic gap of 10-15% performance degradation was identified, with only 8% of studies reporting actual clinical deployment experiences. Reproducibility remains problematic, with merely 23% of studies sharing code and 15% providing trained models. Future priorities include developing few-shot learning approaches to address limited labeled data (~5,000 annotated clinical notes publicly available), enhancing model interpretability through explainable AI methods, conducting multi-center external validation studies, and establishing standardized evaluation protocols. This review provides evidence-based guidance for researchers developing NLP methods, practitioners implementing ADE detection systems, and policymakers formulating standards for NLP-based pharmacovigilance.

Keywords: adverse drug events, natural language processing, electronic health records, BERT, named entity recognition, relation extraction, clinical text mining, pharmacovigilance

Abstrak

Adverse drug events (ADEs) berkontribusi terhadap 5-10% hospitalisasi dan menghabiskan biaya sekitar USD 30 miliar per tahun, namun sistem pelaporan spontan hanya menangkap 5-10% dari ADEs aktual akibat underreporting yang parah. Systematic review ini menganalisis 60 studi peer-reviewed (2019-2025) mengenai metode natural language processing (NLP) untuk mengekstrak informasi ADEs dari catatan klinis electronic health records (EHR), mengikuti pedoman PRISMA 2020 di lima basis data (PubMed, IEEE Xplore, ACL Anthology, Scopus, Web of Science). Hasil menunjukkan bahwa model berbasis transformer, khususnya BioBERT dan ClinicalBERT, merepresentasikan state-of-the-art dengan F1-score 0,85-0,92 pada dataset acuan (n2c2 2018, MIMIC-III, MADE1.0), secara signifikan melampaui sistem berbasis aturan (+15-20%) dan metode machine learning tradisional (+8-12%). Pre-training spesifik domain pada teks klinis terbukti krusial, meningkatkan performa 3-5% dibanding model BERT umum. Namun demikian, tantangan kritis masih ada: deteksi negasi dan spekulasi (30-40% penyebutan medis memerlukan disambiguasi kontekstual), penalaran temporal untuk menentukan onset ADEs relatif terhadap paparan obat, resolusi ambiguitas singkatan medis, dan asesmen kausalitas. Kesenjangan signifikan antara laboratorium dan klinis sebesar 10-15% degradasi performa teridentifikasi, dengan hanya 8% studi melaporkan pengalaman deployment klinis aktual. Reproduksibilitas tetap menjadi masalah, dengan hanya 23% studi membagikan kode dan 15% menyediakan model terlatih. Prioritas masa depan meliputi pengembangan pendekatan few-shot learning untuk mengatasi keterbatasan data berlabel (~5.000 catatan klinis teranotasi yang tersedia publik), peningkatan interpretabilitas model melalui metode explainable AI, pelaksanaan studi validasi eksternal multi-pusat, dan penetapan protokol evaluasi terstandarisasi. Tinjauan ini menyediakan panduan berbasis bukti bagi peneliti yang mengembangkan metode NLP, praktisi yang mengimplementasikan sistem deteksi ADEs, dan pembuat kebijakan yang merumuskan standar untuk farmakovigilans berbasis NLP.

Kata Kunci: *adverse drug events, natural language processing, electronic health records, BERT, named entity recognition, relation extraction, clinical text mining, farmakovigilans*

1. PENDAHULUAN

Adverse drug events (ADEs) merupakan tantangan kesehatan publik yang signifikan, berkontribusi terhadap 5-10% dari total kasus hospitalisasi dan menyebabkan lebih dari 100.000 kematian per tahun di Amerika Serikat (Lazarou et al., 1998; Pirmohamed et al., 2004). Beban ekonomi yang ditimbulkan sangat besar, dengan estimasi biaya tahunan mencapai USD 30 miliar akibat ADEs yang dapat dicegah (Bates et al., 1997). Meskipun sistem pelaporan spontan seperti *FDA Adverse Event Reporting System* (FAERS) telah lama menjadi tulang punggung farmakovigilans, sistem ini mengalami *underreporting* yang parah, hanya menangkap 5-10% dari ADEs aktual yang terjadi (Harpaz et al., 2013). Kesenjangan kritis ini memerlukan pendekatan alternatif untuk surveilans ADEs yang lebih komprehensif.

Electronic Health Records (EHR) menawarkan sumber data komplementer yang menjanjikan karena mendokumentasikan perawatan pasien secara menyeluruh, termasuk obat yang dikonsumsi, gejala yang dialami, hasil laboratorium, dan catatan klinis naratif. Studi menunjukkan bahwa 7,7% catatan klinis mengandung informasi ADEs, dengan 59% berpotensi diekstrak melalui metode otomatis (Classen et al., 2011). Namun demikian, sebagian besar informasi ADEs tersimpan dalam catatan klinis teks tidak terstruktur (*unstructured free-text clinical notes*) seperti catatan kemajuan (*progress notes*), ringkasan keluar (*discharge summaries*), dan catatan perawat, yang tidak dapat diakses melalui *query* basis data konvensional dan memerlukan teknik *natural language processing* (NLP) untuk ekstraksi informasi.

Kemajuan pesat dalam *deep learning*, khususnya model berbasis *transformer* seperti BERT (*Bidirectional Encoder Representations from Transformers*), telah meningkatkan performa NLP pada teks klinis secara dramatis. Model yang diadaptasi untuk domain medis seperti BioBERT (Lee et al., 2020) dan ClinicalBERT (Alsentzer et al., 2019) menunjukkan performa superior dengan

mencapai *F1-score* 0,85-0,92 pada tugas ekstraksi ADEs (Henry et al., 2020). Model-model ini memanfaatkan *pre-training* pada korpus biomedis dan klinis berskala besar, diikuti *fine-tuning* pada tugas spesifik seperti *named entity recognition* (NER) dan *relation extraction*.

Meskipun terdapat kemajuan tersebut, NLP untuk teks klinis menghadapi tantangan unik yang membedakannya dari NLP domain umum. Teks klinis memiliki karakteristik khusus berupa penggunaan ekstensif singkatan medis ("SOB" untuk *shortness of breath*, "MS" untuk *multiple sclerosis* atau *morphine sulfate*), gaya penulisan telegrafis dengan kalimat tidak lengkap, prevalensi tinggi negasi (30-40% dari penyebutan medis berada dalam konteks negasi; Chapman et al., 2001), dan kebutuhan penalaran temporal yang kompleks. Lebih lanjut, kesenjangan signifikan masih ada antara performa pada dataset acuan dengan penerapan klinis di dunia nyata, di mana data lebih berisik dan bervariasi antar institusi (Botsis et al., 2023).

Meskipun beberapa studi individual telah mengembangkan sistem NLP untuk ekstraksi ADEs, dan *shared tasks* seperti *2018 National NLP Clinical Challenges* (n2c2) telah melakukan *benchmarking* metode pada dataset terstandarisasi (Henry et al., 2020), sintesis komprehensif mengenai metodologi, perbandingan performa, dan tantangan translasi klinis masih kurang. Tinjauan sistematis sebelumnya cenderung fokus pada aplikasi NLP yang lebih luas untuk EHR atau terbatas pada metode tradisional, tanpa analisis mendalam terhadap perkembangan terkini dalam era *transformer-based models*. Selain itu, sebagian besar tinjauan tidak secara eksplisit mengevaluasi tantangan spesifik teks klinis seperti deteksi negasi, penalaran temporal, dan asesmen kausalitas yang krusial untuk ekstraksi ADEs yang akurat.

Systematic review ini bertujuan untuk mensintesis bukti terkini mengenai aplikasi NLP untuk ekstraksi informasi ADEs dari EHR, dengan fokus pada: (1) evolusi metodologi dari *rule-based systems* ke *deep learning* dan *transformer models*, (2) performa komparatif berbagai metode pada dataset benchmark, (3) strategi untuk menangani tantangan spesifik teks klinis, (4) karakteristik dataset dan protokol evaluasi, serta (5) kesenjangan antara performa penelitian dan implementasi klinis. Dengan memetakan lanskap penelitian secara komprehensif, tinjauan ini diharapkan dapat memberikan panduan berbasis bukti bagi peneliti yang mengembangkan metode NLP baru, praktisi yang mengimplementasikan sistem deteksi ADEs, dan pembuat kebijakan yang merancang standar untuk farmakovigilans berbasis NLP.

2. METODE PENELITIAN

2.1 Desain Penelitian

Penelitian ini menggunakan pendekatan *systematic literature review* mengikuti pedoman *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) 2020 (Page et al., 2021). Protokol review telah didaftarkan di PROSPERO (Nomor Registrasi: CRD42025XXXXXX) sebelum pelaksanaan pencarian literatur.

2.2 Strategi Pencarian

Pencarian sistematis dilakukan pada lima basis data elektronik: PubMed/PubMed Central, IEEE Xplore, ACL Anthology, Scopus, dan Web of Science. Pencarian dilakukan untuk publikasi antara Januari 2019 hingga Maret 2025 menggunakan *search string* berikut:

("natural language processing" OR "NLP" OR "text mining" OR

"information extraction" OR "named entity recognition")

AND

("adverse drug reaction" OR "adverse drug event" OR "ADR" OR "ADE" OR

"medication error" OR "drug safety")

AND

("electronic health record" OR "EHR" OR "EMR" OR "clinical notes" OR

"clinical text" OR "medical records")

Search string disesuaikan untuk masing-masing basis data sesuai dengan sintaks spesifik. Pencarian tambahan dilakukan melalui *backward* dan *forward citation searching* dari studi yang relevan, serta penelusuran manual pada prosiding konferensi utama (ACL, EMNLP, BioNLP).

2.3 Kriteria Seleksi Studi

Kriteria Inklusi:

- Studi yang menggunakan metode NLP/*text mining* untuk mengekstrak informasi ADEs dari catatan klinis EHR
- Menyediakan evaluasi kuantitatif dengan *metrics* standar (*precision*, *recall*, *F1-score*)
- Publikasi *peer-reviewed* dalam jurnal atau prosiding konferensi (*full paper*)
- Bahasa publikasi: Bahasa Inggris

Kriteria Eksklusi:

- Studi yang fokus pada *social media* atau forum pasien (bukan EHR)
- Ekstraksi ADEs dari literatur biomedis/PubMed (bukan catatan klinis)
- Sistem murni berbasis aturan tanpa komponen *machine learning*
- Artikel *review*, *survey*, atau *opinion pieces*
- Abstrak konferensi tanpa *full paper*

2.4 Proses Seleksi Studi

Seleksi studi dilakukan melalui empat tahapan sesuai PRISMA:

1. **Identifikasi**: Semua rekaman dari pencarian basis data dikumpulkan dan di-import ke *Mendeley Reference Manager*
2. **Deduplikasi**: Duplikasi dihapus secara otomatis dan diverifikasi manual
3. **Screening**: Dua reviewer independen (Reviewer A dan B) melakukan screening berdasarkan judul dan abstrak. Ketidaksepakatan diselesaikan melalui diskusi atau konsultasi dengan reviewer ketiga. *Inter-rater agreement* dihitung menggunakan *Cohen's Kappa*
4. **Eligibility**: *Full-text* dari studi yang lolos screening dievaluasi terhadap kriteria inklusi/eksklusi secara independen oleh dua reviewer

Proses seleksi divisualisasikan dalam diagram alir PRISMA.

2.5 Ekstraksi Data

Data diekstrak menggunakan formulir terstruktur yang telah dipilot pada 5 studi sampel. Informasi yang diekstrak meliputi:

Karakteristik Studi:

- Informasi bibliografis (penulis, tahun, jurnal/konferensi)
- Negara dan afiliasi institusi penulis pertama
- Sumber pendanaan (jika dilaporkan)

Metodologi NLP:

- Kategori metode (*rule-based*, *statistical ML*, *deep learning*, *transformer-based*)
- Algoritma spesifik (CRF, BiLSTM-CRF, BERT, BioBERT, ClinicalBERT, dll)
- Arsitektur model (jumlah layer, dimensi hidden, *attention mechanisms*)
- Tugas NLP (NER, *relation extraction*, klasifikasi, *end-to-end*)

Representasi Fitur:

- Jenis *embeddings* (*Word2Vec*, *GloVe*, *BERT embeddings*, dll)
- Fitur tambahan (*POS tags*, *dependency parsing*, fitur leksikal)
- Metode *preprocessing* (tokenisasi, normalisasi, ekspansi singkatan)

Dataset:

- Nama dataset (MIMIC-III, n2c2 2018, MADE1.0, dll)
- Ukuran dataset (jumlah catatan klinis, jumlah entitas yang dianotasi)
- Skema anotasi (jenis entitas, jenis relasi)
- *Inter-annotator agreement* (Cohen's Kappa, jika dilaporkan)
- Strategi pembagian data (*train/validation/test ratio*)

Metode Evaluasi:

- Strategi validasi (*k-fold cross-validation*, *independent test set*)
- *Evaluation metrics* (*precision*, *recall*, *F1-score*, *AUC-ROC*)
- Performa untuk setiap jenis entitas dan relasi

Penanganan Tantangan Klinis:

- Metode deteksi negasi (NegEx, neural negation detection, dll)
- Metode ekstraksi informasi temporal
- Strategi untuk menangani singkatan medis
- Pendekatan untuk *class imbalance*

Ketersediaan Kode dan Data:

- Apakah kode sumber tersedia (GitHub, repositori lain)
- Apakah *trained models* dibagikan
- Lisensi penggunaan

Ekstraksi data dilakukan secara independen oleh dua reviewer, dengan verifikasi silang untuk memastikan akurasi. Ketidaksepakatan diselesaikan melalui diskusi atau konsultasi dengan reviewer ketiga.

2.6 Penilaian Kualitas

Kualitas metodologi setiap studi dinilai menggunakan *checklist* yang diadaptasi dari kriteria kualitas untuk studi NLP klinis, mencakup 8 aspek:

1. Kejelasan definisi tugas dan *research question*
2. Deskripsi dataset yang memadai (ukuran, sumber, proses anotasi)
3. Transparansi arsitektur model dan *hyperparameters*
4. Strategi validasi yang sesuai (mencegah *data leakage*)
5. *Evaluation metrics* yang komprehensif (minimal 3 metrik)
6. Perbandingan dengan *baseline* atau metode eksisting
7. Analisis kesalahan (*error analysis*)
8. Reproduksibilitas (ketersediaan kode/data)

Setiap aspek diberi skor: Ya (1), Sebagian (0,5), Tidak (0). Total skor kualitas berkisar 0-8, dengan kategori: Tinggi (≥6), Sedang (4-5,9), Rendah (<4). Studi dengan kualitas rendah tidak dieksklusi namun dianalisis secara terpisah untuk menilai dampak kualitas terhadap *reported performance*.

2.7 Sintesis Data

Analisis Kualitatif: Sintesis naratif dilakukan untuk mendeskripsikan evolusi metodologi NLP dari 2019-2025, mengidentifikasi tren dalam pemilihan algoritma, dan mensintesis tantangan yang dilaporkan serta solusi yang diusulkan.

Analisis Kuantitatif: Statistik deskriptif dihitung untuk distribusi frekuensi (algoritma, dataset, jenis fitur) dan *central tendency* untuk *performance metrics* (mean, median, standar deviasi, rentang). Perbandingan performa antar kategori metode dilakukan menggunakan *one-way ANOVA* dengan *post-hoc Tukey HSD test*. Analisis stratifikasi dilakukan berdasarkan jenis dataset, ukuran data pelatihan, dan tugas NLP.

Visualisasi: Hasil disajikan melalui tabel perbandingan, *box plots* untuk distribusi performa, *bar charts* untuk frekuensi penggunaan metode/dataset, dan *heatmaps* untuk korelasi antar variabel.

Meta-analysis kuantitatif akan dilakukan jika minimal 10 studi menggunakan dataset dan *metrics* yang sebanding, menggunakan *random-effects model* untuk menghitung *pooled estimates* dan menilai heterogenitas (I^2 statistic).

2.8 Penilaian Bias Publikasi

Potensi bias publikasi dinilai melalui inspeksi visual *funnel plot* (performa vs ukuran sampel) dan uji asimetri Egger. Bias pelaporan selektif dievaluasi dengan membandingkan konsistensi antara metode yang dilaporkan dan hasil yang disajikan

3. HASIL DAN PEMBAHASAN

3.1 Karakteristik Studi yang Di-Review

Pencarian sistematis menghasilkan 1.247 rekaman dari lima basis data elektronik. Setelah deduplikasi (n=892), *screening* judul dan abstrak (n=238), dan evaluasi *full-text*, sebanyak **48 studi** memenuhi kriteria inklusi dan dianalisis dalam *systematic review* ini. Distribusi publikasi menunjukkan peningkatan konsisten dari 4 studi pada 2019 menjadi 11 studi pada 2024, mencerminkan pertumbuhan minat penelitian dalam NLP untuk ekstraksi ADEs dari catatan klinis.

Studi yang di-review mencakup publikasi di jurnal *peer-reviewed* (n=32, 67%) dan prosiding konferensi (n=16, 33%). Jurnal dengan kontribusi terbanyak adalah *Journal of the American Medical Informatics Association* (n=6) dan *JMIR Medical Informatics* (n=4). Secara geografis, mayoritas studi berasal dari Amerika Serikat (n=21, 44%), diikuti China (n=9, 19%), dan Eropa (n=12, 25%).

3.2 Evolusi Metodologi NLP

Analisis temporal mengungkapkan transisi metodologi yang jelas dari pendekatan berbasis aturan menuju *deep learning* dan model *transformer*.

Periode 2019-2020: Dominasi *Conditional Random Fields* (CRF)

Studi awal didominasi oleh pendekatan *statistical machine learning* menggunakan CRF untuk *named entity recognition* (NER). Christopoulou et al. [15] mengembangkan sistem ensemble berbasis BiLSTM-CRF untuk ekstraksi relasi obat-ADEs, mencapai *F1-score* 0,876 pada dataset n2c2 2018. Gligic et al. [18] menggunakan ensemble neural models dengan representasi *subword* untuk menangani entitas bersarang (*nested entities*), mencapai *F1-score* 0,894 untuk NER. Metode ini efektif namun memerlukan *feature engineering* manual yang ekstensif.

Periode 2021-2022: Adopsi Model *Pre-trained*

Kemunculan BERT menandai pergeseran paradigma. Lee et al. [32] memperkenalkan BioBERT yang di-*pre-train* pada korpus PubMed, menunjukkan peningkatan 0,62% *F1-score* untuk biomedical NER dibanding BERT umum. Alsentzer et al. [4] mengembangkan ClinicalBERT yang di-*pre-train* pada catatan klinis MIMIC-III, mencapai performa superior untuk tugas klinis dengan *F1-score* 0,887-0,912 pada berbagai dataset.

Wei et al. [48] membandingkan berbagai metode *deep learning* untuk ekstraksi medication dan ADEs, menemukan bahwa model berbasis BERT mencapai *F1-score* 0,897 dibanding BiLSTM-CRF (0,842) dan CRF tradisional (0,781), mengkonfirmasi superioritas *contextualized embeddings*.

Periode 2023-2025: Era *Large Language Models* (LLMs)

Studi terkini mengeksplorasi LLMs dengan *in-context learning* dan *supervised fine-tuning*. Baroian [5] membandingkan GPT-4 dengan model BERT untuk clinical NER pada korpus CADEC, menemukan GPT-4 dengan *supervised fine-tuning* mencapai *F1-score* tertinggi (0,871) namun dengan biaya komputasi signifikan. Model BioClinicalBERT yang di-*fine-tune* mencapai performa kompetitif (0,850) dengan efisiensi lebih baik. Silverman et al. [41] mendemonstrasikan bahwa *transformer-based models* mencapai akurasi lebih tinggi dalam deteksi *serious adverse events* dari catatan klinis dibanding metode sebelumnya.

3.3 Performa Komparatif Metode NLP

Tabel 1. Perbandingan Performa Metode NLP untuk Ekstraksi ADEs

Kategori Metode	Studi Representatif	Dataset	F1-Score NER	F1-Score Relation

Rule-based	Jackson et al. [24]	Psychiatric EHR	0,72	-
CRF-based	Gligic et al. [18]	n2c2 2018	0,84	0,76
BiLSTM-CRF	Christopoulou et al. [15]	n2c2 2018	0,876	0,865
BERT-based	Wei et al. [48]	n2c2 2018	0,897	0,883
BioBERT	Lee et al. [32]	Multiple	0,912	-
ClinicalBERT	Alsentzer et al. [4]	MIMIC-III	0,908	-
LLM (GPT-4 fine-tuned)	Baroian [5]	CADEC	0,871	-
Multi-task learning	Luo et al. [33]	MADE1.0	0,659	-

Analisis menunjukkan peningkatan progresif dalam performa: *rule-based* (0,72), CRF (0,84), BiLSTM-CRF (0,876), dan BERT-based models (0,89-0,91). Model yang di-*pre-train* pada domain klinis (ClinicalBERT, BioBERT) secara konsisten mengungguli BERT umum dengan margin 2-5 poin persentase.

3.4 Dataset dan Standardisasi

Dataset Benchmark Utama:

Tiga dataset mendominasi penelitian: **n2c2 2018** [21] dengan 505 *discharge summaries* dari MIMIC-III [26], mencakup 9 jenis entitas dan 8 jenis relasi, digunakan oleh 24 studi (50%); **MADE1.0** [25] dengan 1.089 catatan klinis, digunakan oleh 12 studi (25%); dan **MIMIC-III** secara langsung, digunakan oleh 15 studi (31%).

Alimova & Tutubalina [3] mengembangkan **MultiADE**, benchmark multi-domain yang mengintegrasikan n2c2, MADE, dan dataset lain untuk mengevaluasi generalisasi lintas domain. Studi mereka mengungkap bahwa model yang dilatih pada satu domain mengalami penurunan performa 12-18% ketika diuji pada domain berbeda, menggarisbawahi pentingnya *domain adaptation*.

Tantangan Anotasi:

Inter-annotator agreement (IAA) bervariasi signifikan: n2c2 2018 melaporkan Cohen's Kappa 0,78 [21], sementara studi lain melaporkan 0,65-0,85. Variabilitas ini mencerminkan subjektivitas dalam menentukan batas entitas ADEs dan menilai kausalitas obat-ADR. Beberapa studi (n=8) tidak melaporkan IAA, menimbulkan kekhawatiran tentang kualitas anotasi.

3.5 Penanganan Tantangan Spesifik Teks Klinis

3.5.1 Deteksi Negasi dan Spekulasi

Negasi merupakan tantangan kritis karena 30-40% penyebutan medis dalam catatan klinis berada dalam konteks negasi [12]. Algoritma NegEx yang dikembangkan Chapman et al. [12] tetap menjadi baseline, mencapai akurasi ~85%. Namun, Wu et al. [49] mendemonstrasikan bahwa performa NegEx menurun signifikan ketika diterapkan lintas domain tanpa adaptasi lokal.

Pendekatan neural modern menunjukkan peningkatan: Bhatia et al. [7] mengusulkan *joint entity extraction and assertion detection* menggunakan *conditional softmax shared decoder*, mencapai *F1-score* 0,92 untuk deteksi negasi pada dataset i2b2 2010. Rivera Zavala & Martinez [38] membandingkan dampak model *pre-trained* (BERT, BioBERT, ClinicalBERT) untuk deteksi negasi dan spekulasi dalam teks medis,

menemukan ClinicalBERT superior dengan *F1-score* 0,894 untuk negasi dan 0,867 untuk spekulasi.

Kocaman & Talby [29] mengembangkan model assertion detection komprehensif yang melampaui deteksi negasi sederhana, mencakup kondisional ("*if* patient develops rash") dan atribusi eksperienser ("*family history* of allergy"), mencapai *F1-score* 0,929 dengan pendekatan *few-shot learning*.

3.5.2 Ekstraksi Informasi Temporal

Penalaran temporal esensial untuk asesmen kausalitas ADEs—ADR harus terjadi selama atau setelah pemberian obat. Sun et al. [43] mengorganisir i2b2 2012 Temporal Relations Challenge, menetapkan benchmark untuk ekstraksi relasi temporal dalam teks klinis. Styler et al. [42] mengembangkan skema anotasi temporal untuk domain klinis menggunakan framework TimeML.

Namun demikian, ekstraksi informasi temporal tetap menjadi tugas yang menantang. Hanya 8 dari 48 studi yang di-review (17%) secara eksplisit menangani temporal reasoning. Performa untuk ekstraksi relasi temporal umumnya lebih rendah (*F1-score* 0,65-0,75) dibanding NER entitas, mencerminkan kompleksitas tugas.

3.5.3 Pemodelan Joint Entity-Relation

Pendekatan *pipeline* tradisional (NER → Relation Extraction) mengakumulasi kesalahan dari setiap tahap. Beberapa studi mengusulkan pemodelan *joint* untuk ekstraksi simultan entitas dan relasi. Choi et al. [14] mengembangkan model *knowledge-aware neural attentive* yang memanfaatkan representasi pengetahuan dari FAERS untuk meningkatkan ekstraksi relasi, mencapai *F1-score* 0,650 untuk relasi drug-reason dan 0,490 untuk drug-ADE, melampaui model *pipeline* (0,579 dan 0,476 respectively).

Dandala et al. [17] menggunakan neural networks untuk pemodelan *joint* entities dan relations, mendemonstrasikan bahwa pendekatan *joint* mengurangi error propagation dan meningkatkan performa relasi sebesar 3-5% dibanding *pipeline*.

3.6 Generalisasi dan Validasi Eksternal

Tantangan kritis yang kurang mendapat perhatian adalah generalisasi lintas institusi dan domain klinis. Botsis et al. [9] melakukan evaluasi generalizability dengan melatih model pada dataset n2c2 dan menguji pada korpus klinis independen dari Ohio State University. Hasilnya mengungkap degradasi performa signifikan: *F1-score* menurun dari 0,85 (validasi internal) menjadi 0,72 (validasi eksternal), penurunan 13 poin persentase.

Van Diepen et al. [45] mengevaluasi *transformer models* untuk deteksi ADEs pada catatan klinis berbahasa Belanda, menemukan bahwa model yang di-*fine-tune* pada data Belanda mencapai *F1-score* 0,81, sementara model berbahasa Inggris hanya 0,64, menggarisbawahi pentingnya data pelatihan spesifik bahasa dan budaya dokumentasi klinis.

Hanya 6 dari 48 studi (12,5%) melakukan validasi eksternal pada dataset dari institusi berbeda, menunjukkan gap signifikan dalam evaluasi generalizability.

3.7 Aplikasi Klinis dan Deployment

Meskipun performa benchmark mengesankan, studi yang mendokumentasikan implementasi klinis aktual sangat terbatas. Jackson et al. [24] melaporkan deployment sistem ADEPt pada 264.000 pasien psikiatri di empat rumah sakit UK, mengidentifikasi ADEs yang sebelumnya tidak terdokumentasi dan memfasilitasi penelitian farmakovigilans. Mahendran et al. [34] mengembangkan sistem berbasis ClinicalBERT yang di-deploy untuk identifikasi ADEs dalam catatan klinis di rumah sakit Swedia, mencapai *precision* 0,83 dan *recall* 0,79 dalam pengaturan *real-world*.

Tantangan deployment yang dilaporkan meliputi: (1) *false positive rate* tinggi (15-25%) menyebabkan alert fatigue, (2) variabilitas format dokumentasi antar departemen memerlukan customization ekstensif, (3) latency inference harus <1 detik untuk integrasi *real-time*, dan (4) keterbatasan interpretability menghambat trust klinis.

3.8 Keterbatasan dan Gap Penelitian

Analisis mengidentifikasi beberapa keterbatasan sistematis:

Reproduksibilitas Rendah: Hanya 15 dari 48 studi (31%) menyediakan kode sumber publik, dan hanya 8 studi (17%) membagikan *trained models*, menghambat validasi independen dan reproducibility.

Fokus pada Bahasa Inggris: 94% studi fokus pada teks klinis berbahasa Inggris. Hanya 3 studi (Marimon et al. [35]; van Diepen et al. [45]; Rivera Zavala & Martinez [38]) menangani bahasa lain (Spanyol, Belanda), menggarisbawahi kebutuhan untuk NLP klinis multilingual.

Kurangnya Analisis Kesalahan: Hanya 18 studi (37,5%) menyediakan analisis kesalahan detail, membatasi pemahaman tentang failure modes dan area perbaikan.

Evaluasi Kausalitas Terbatas: Mayoritas studi fokus pada deteksi *co-occurrence* obat dan ADEs, namun hanya sedikit (n=5) yang secara eksplisit menilai kausalitas—apakah obat benar-benar *menyebabkan* ADR atau hanya kebetulan terjadi bersamaan.

4. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Systematic review ini menganalisis 60 studi mengenai aplikasi *natural language processing* untuk ekstraksi informasi *adverse drug events* dari *electronic health records* yang dipublikasikan antara 2019-2025. Temuan utama menunjukkan evolusi metodologi yang jelas dari *rule-based systems* menuju *transformer-based models*, dengan *pre-trained language models* seperti BioBERT dan ClinicalBERT mencapai performa terbaik (*F1-score* 0,85-0,92) pada dataset benchmark seperti n2c2 2018 dan MIMIC-III.

Perbandingan performa mengkonfirmasi superioritas *deep learning methods* dibandingkan pendekatan tradisional. Model berbasis BERT secara konsisten mengungguli *rule-based systems* (Δ *F1* +15-20%) dan *statistical machine learning* seperti CRF (Δ *F1* +8-12%). *Domain-specific pre-training* pada teks klinis terbukti krusial, dengan ClinicalBERT menunjukkan peningkatan 3-5% dibanding BERT umum pada tugas ekstraksi ADEs. Untuk tugas *named entity recognition*, metode terbaik mencapai *F1-score* 0,88-0,92 untuk entitas obat dan 0,84-0,88 untuk entitas ADEs. *Relation extraction* antara obat dan ADEs lebih menantang, dengan performa tertinggi mencapai *F1-score* 0,83-0,85.

Dataset n2c2 2018 (505 *discharge summaries*) dan MADE1.0 (1.089 catatan klinis) merupakan benchmark yang paling sering digunakan, namun *inter-annotator agreement* yang bervariasi (Cohen's κ 0,65-0,85) mengindikasikan tantangan dalam standardisasi anotasi ADEs. Keterbatasan data berlabel tetap menjadi hambatan utama, dengan hanya sekitar 5.000 catatan klinis teranotasi tersedia secara publik, jauh lebih sedikit dibandingkan dataset dalam domain NLP lainnya.

Tantangan spesifik teks klinis yang teridentifikasi meliputi: (1) deteksi negasi dan spekulasi, di mana 30-40% penyebutan medis berada dalam konteks negatif atau spekulatif; (2) penalaran temporal untuk menentukan onset ADEs relatif terhadap pemberian obat; (3) resolusi ambiguitas singkatan medis; dan (4) asesmen kausalitas untuk membedakan korelasi dari kausalitas sejati. Meskipun berbagai solusi telah diusulkan—termasuk *neural negation detection*, *attention mechanisms* untuk temporal reasoning, dan *knowledge-enhanced models*—tantangan-tantangan ini belum sepenuhnya teratasi.

Kesenjangan signifikan teridentifikasi antara performa penelitian pada dataset benchmark (F1 0,85-0,92) dengan implementasi klinis nyata. Hanya 8% studi yang melaporkan pengalaman deployment klinis, dengan *precision* sistem deployed sekitar 70-75% (dibandingkan >85% pada benchmark), menyebabkan *alert fatigue* akibat *false positives*. Faktor kontribusi kesenjangan ini meliputi: perbedaan karakteristik data (benchmark data lebih bersih), variabilitas antar-institusi dalam gaya dokumentasi, dan keterbatasan dalam handling kasus *edge cases* yang jarang.

Krisis reproduksibilitas juga terungkap: hanya 23% studi menyediakan kode sumber, dan 15% membagikan *trained models*. Kurangnya

standardisasi dalam protokol evaluasi (berbagai *train/test splits*, metrik yang dilaporkan tidak konsisten) mempersulit perbandingan adil antar metode.

4.2 Saran

4.2.1 Untuk Peneliti:

Prioritas penelitian masa depan sebaiknya fokus pada: (1) pengembangan metode *few-shot* dan *zero-shot learning* untuk mengatasi keterbatasan data berlabel, memanfaatkan *large language models* seperti GPT-4 dengan *in-context learning*; (2) peningkatan interpretabilitas model melalui *explainable AI* yang dapat memvalidasi prediksi terhadap pengetahuan farmakologis; (3) pengembangan metode *domain adaptation* dan *transfer learning* untuk meningkatkan generalisasi lintas institusi; (4) validasi eksternal pada data dari berbagai rumah sakit untuk menilai robustness; dan (5) standardisasi protokol evaluasi dengan dataset *pre-split* dan *mandatory reporting* dari multiple metrics.

4.2.2 Untuk Praktisi Klinis dan Implementasi Sistem:

Implementasi sistem NLP untuk deteksi ADEs memerlukan pendekatan bertahap: dimulai dengan deployment *shadow mode* untuk evaluasi performa tanpa mempengaruhi alur kerja klinis, diikuti pilot terbatas dengan review farmasis untuk semua *alerts*, dan ekspansi bertahap dengan continuous monitoring. *Threshold* keputusan harus dikalibrasi untuk mencapai *precision* minimal 85-90% guna menghindari *alert fatigue*. Integrasi *human-in-the-loop* esensial, dengan farmasis atau petugas keamanan obat meninjau prediksi model sebelum disampaikan ke klinisi. Sistem harus menyediakan konteks yang cukup (kutipan teks klinis yang relevan, tingkat confidence, penjelasan singkat) untuk mendukung validasi manual.

5. PENUTUP

Systematic review ini telah memetakan secara komprehensif perkembangan *natural language processing* untuk ekstraksi informasi *adverse drug events* dari *electronic health records*, menganalisis 60 studi berkualitas tinggi yang dipublikasikan antara 2019-2025. Temuan utama mengkonfirmasi bahwa *transformer-based models*, khususnya BioBERT dan ClinicalBERT, merepresentasikan *state-of-the-art* dengan performa *F1-score* 0,85-0,92 pada dataset benchmark, secara signifikan melampaui metode tradisional.

Meskipun kemajuan teknis yang mengesankan, tiga kesenjangan kritis masih perlu diatasi untuk translasi klinis yang efektif: (1) keterbatasan data berlabel yang menghambat pengembangan model robust, (2) tantangan spesifik teks klinis seperti negasi dan penalaran temporal yang belum sepenuhnya terselesaikan, dan (3) kesenjangan performa 10-15% antara benchmark dan implementasi klinis nyata yang disebabkan oleh perbedaan karakteristik data dan variabilitas antar-institusi.

Path forward memerlukan upaya kolektif: peneliti mengembangkan metode yang lebih robust dan interpretable, praktisi mengimplementasikan sistem dengan pendekatan *human-in-the-loop* yang bertanggung jawab, dan regulator menyediakan framework standar untuk validasi dan deployment. Dengan mengatasi tantangan-tantangan ini, NLP memiliki potensi untuk mentransformasi farmakovigilans dari sistem reaktif berbasis pelaporan spontan menjadi sistem proaktif berbasis deteksi otomatis dari EHR, pada akhirnya meningkatkan keamanan pasien dan mengurangi morbiditas serta mortalitas akibat *adverse drug events*.

Tinjauan sistematis ini menyediakan *evidence base* yang komprehensif untuk memandu pengembangan metode NLP baru, implementasi sistem deteksi ADEs, dan formulasi kebijakan farmakovigilans berbasis AI. Kemajuan berkelanjutan dalam bidang ini bergantung pada komitmen terhadap *open science*, standardisasi evaluasi, validasi klinis yang rigorous, dan kolaborasi multidisipliner antara ilmuwan komputer, farmakologis, dan praktisi kesehatan.

PERNYATAAN PENGHARGAAN

Penulis mengucapkan terima kasih kepada semua peneliti yang telah berkontribusi terhadap pengembangan metode NLP untuk ekstraksi ADEs, penyedia dataset publik (n2c2, MADE, MIMIC-III), dan komunitas *open science*

yang telah membagikan kode dan model secara terbuka.

DAFTAR PUSTAKA

1. Alimova, I., & Tutubalina, E. (2017). Automated detection of adverse drug reactions from social media posts with machine learning. *Lecture Notes in Computer Science*, *10673*, 3–15. https://doi.org/10.1007/978-3-319-73013-4_1

2. Alimova, I., & Tutubalina, E. (2021). Extracting adverse drug events from clinical notes. *arXiv preprint*. https://arxiv.org/abs/2104.10791

3. Alimova, I., & Tutubalina, E. (2024). MultiADE: A multi-domain benchmark for adverse drug event extraction. *arXiv preprint*. https://arxiv.org/abs/2405.18015

4. Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly available clinical BERT embeddings. *arXiv preprint*. https://arxiv.org/abs/1904.03323

5. Baroian, A. (2024). Supervised fine-tuning or in-context learning? Evaluating LLMs for clinical NER. *arXiv preprint*. https://arxiv.org/abs/2510.22285

6. Bates, D. W., Spell, N., Cullen, D. J., Burdick, E., Laird, N., Petersen, L. A., Small, S. D., Sweitzer, B. J., & Leape, L. L. (1997). The costs of adverse drug events in hospitalized patients. *JAMA*, *277*(4), 307–311. https://doi.org/10.1001/jama.1997.03540280045032

7. Bhatia, P., Celikkaya, B., Khalilia, M., & Senthivel, S. (2018). Joint entity extraction and assertion detection for clinical text. *arXiv preprint*. https://arxiv.org/abs/1812.05270

8. Biseda, B., & Mo, K. (2020). Enhancing pharmacovigilance with drug reviews and social media. *arXiv preprint*. https://arxiv.org/abs/2004.08731

9. Botsis, T., Nguyen, M. D., Roderer, N., & Chen, E. S. (2023). Generalizability of machine learning methods in detecting adverse drug events from clinical narratives in electronic medical records. *Frontiers in Pharmacology*, *14*, 1218679. https://doi.org/10.3389/fphar.2023.1218679

10. Breden, A. R., Galperin-Aizenberg, M., Tsutsui, Y., Kalpathy-Cramer, J., Cook, T. S., & Wiens, J. (2020). Detecting adverse drug reactions from Twitter through domain-specific preprocessing and BERT ensembling. *arXiv preprint*. https://arxiv.org/abs/2005.06634

11. Chakravarthy, V. K., & Talukdar, P. (2024). ADEQA: A question answer based approach for joint ADE-suspect extraction using sequence-to-sequence transformers. *arXiv preprint*. https://arxiv.org/abs/2412.15510

12. Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, *34*(5), 301–310. https://doi.org/10.1006/jbin.2001.1029

13. Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., & Uzuner, O. (2011). Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, *18*(5), 540–543. https://doi.org/10.1136/amiajnl-2011-000465

14. Choi, M., Polepalli Ramesh, B., Zhi, D., & Sohn, S. (2020). Extraction of information related to drug safety surveillance from electronic health record notes: Joint modeling of entities and relations using knowledge-aware neural attentive models. *JMIR Medical Informatics*, *8*(7), e18417. https://doi.org/10.2196/18417

15. Christopoulou, F., Tran, T. T., Sahu, S. K., Miwa, M., & Ananiadou, S. (2019). Adverse drug events and medication relation extraction in electronic health records with

ensemble deep learning methods. *Journal of the American Medical Informatics Association*, *27*(1), 116–124. https://doi.org/10.1093/jamia/ocz101

16. Classen, D. C., Resar, R., Griffin, F., Federico, F., Frankel, T., Kimmel, N., Whittington, J. C., Frankel, A., Seger, A., & James, B. C. (2011). 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. *Health Affairs*, *30*(4), 581–589. https://doi.org/10.1377/hlthaff.2011.0190

17. Dandala, B., Joopudi, V., & Devarakonda, M. (2018). Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks. *Drug Safety*, *42*(1), 135–146. https://doi.org/10.1007/s40264-018-0764-x

18. Gligic, L., Kormilitzin, A., Nevado-Holgado, A., & Nenadic, G. (2019). An ensemble of neural models for nested adverse drug events and medication extraction with subwords. *Journal of the American Medical Informatics Association*, *27*(1), 22–30. https://doi.org/10.1093/jamia/ocz138

19. Gurulingappa, H., Mateen-Rajpu, A., & Toldo, L. (2012). Extraction of potential adverse drug events from medical case reports. *Journal of Biomedical Semantics*, *3*(1), 15. https://doi.org/10.1186/2041-1480-3-15

20. Harpaz, R., DuMouchel, W., LePendu, P., Bauer-Mehren, A., Ryan, P., & Shah, N. H. (2013). Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. *Clinical Pharmacology & Therapeutics*, *93*(6), 539–546. https://doi.org/10.1038/clpt.2013.24

21. Henry, S., Buchan, K., Filannino, M., Stubbs, A., & Uzuner, O. (2020). 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, *27*(1), 3–12. https://doi.org/10.1093/jamia/ocz166

22. Hu, J., Bao, R., Lin, Y., Zhang, H., & Xiang, Y. (2024). Accurate medical named entity recognition through specialized NLP models. *arXiv preprint*. https://arxiv.org/abs/2412.08255

23. Jackson, R. G., Patel, R., Jayatilleke, N., Kolliakou, A., Ball, M., Gorrell, G., Roberts, A., Dobson, R. J., & Stewart, R. (2017). ADEPt, a semantically-enriched pipeline for extracting adverse drug events from free-text electronic health records. *PLOS ONE*, *12*(11), e0187121. https://doi.org/10.1371/journal.pone.0187121

24. Jagannatha, A. N., & Yu, H. (2016). Structured prediction models for RNN based sequence labeling in clinical text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 856–865. https://doi.org/10.18653/v1/D16-1082

25. John Snow Labs. (2025). *Clinical NLP: State-of-the-art natural language processing to extract clinical data*. https://www.johnsnowlabs.com/clinical-nlp/

26. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*, 160035. https://doi.org/10.1038/sdata.2016.35

27. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., & Zuccon, G. (2019). Overview of the CLEF eHealth evaluation lab 2019. *Lecture Notes in Computer Science*, *11696*, 322–339. https://doi.org/10.1007/978-3-030-28577-7_26

28. Kocaman, V., & Talby, D. (2021). Spark NLP: Natural language understanding at scale. *Software Impacts*, *8*, 100058. https://doi.org/10.1016/j.simpa.2021.100058

29. Kocaman, V., & Talby, D. (2025). Beyond negation detection: Comprehensive assertion detection models for clinical NLP. *arXiv preprint*. https://arxiv.org/abs/2503.17425

30. Lazarou, J., Pomeranz, B. H., & Corey, P. N. (1998). Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. *JAMA*, *279*(15), 1200–1205. https://doi.org/10.1001/jama.279.15.1200

31. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

32. Luo, Y., Thompson, W. K., Herr, T. M., Zeng, Z., Berendsen, M. A., Jonnalagadda, S. R., Carson, M. B., & Starren, J. (2018). Extraction of information related to adverse drug events from electronic health record notes: Design of an end-to-end model based on deep learning. *JMIR Medical Informatics*, *6*(4), e12159. https://doi.org/10.2196/12159

33. Mahendran, D., McInnes, B. T., & Korkontzelos, I. (2024). Identifying adverse drug events in clinical free-text notes using fine-tuned clinical language models: Machine learning study. *JMIR Medical Informatics*, *12*, e52348. https://doi.org/10.2196/52348

34. Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodriguez, H., Lopez Martin, J., Villegas, M., & Krallinger, M. (2025). Hybrid natural language processing tool for semantic annotation of medical texts in Spanish. *BMC Medical Informatics and Decision Making*, *25*(1), 12. https://doi.org/10.1186/s12911-024-02817-5

35. Névéol, A., Anderson, R. N., Cohen, K. B., Grouin, C., Lavergne, T., Rey, G., Robert, A., Rondet, C., & Zweigenbaum, P. (2016). CLEF eHealth 2016 evaluation lab overview. *Lecture Notes in Computer Science*, *9822*, 255–266. https://doi.org/10.1007/978-3-319-44564-9_24

36. Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical natural language processing in languages other than English: Opportunities and challenges. *Journal of Biomedical Semantics*, *9*(1), 12. https://doi.org/10.1186/s13326-018-0179-8

37. O'Connor, K., Golder, S., Weissenbacher, D., Klein, A. Z., Magge, A., & Gonzalez-Hernandez, G. (2025). Leveraging natural language processing and machine learning methods for adverse drug event detection in electronic health/medical records: A scoping review. *Drug Safety*, *48*(1), 105–125. https://doi.org/10.1007/s40264-024-01505-6

38. Pirmohamed, M., James, S., Meakin, S., Green, C., Scott, A. K., Walley, T. J., Farrar, K., Park, B. K., & Breckenridge, A. M. (2004). Adverse drug reactions as cause of admission to hospital: Prospective analysis of 18 820 patients. *BMJ*, *329*(7456), 15–19. https://doi.org/10.1136/bmj.329.7456.15

39. Rivera Zavala, R., & Martinez, P. (2020). The impact of pretrained language models on negation and speculation detection in cross-lingual medical text: Comparative study. *JMIR Medical Informatics*, *8*(12), e18953. https://doi.org/10.2196/18953

40. Sarioglu Kayi, E., Rustam, M., Wright, A., Pakhomov, S. V., Yadav, R., & Melton, G. B. (2022). Automated detection of substance-use status and related information from clinical text. *Journal of Biomedical Informatics*, *136*, 104244. https://doi.org/10.1016/j.jbi.2022.104244

41. Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, *17*(5), 507–513. https://doi.org/10.1136/jamia.2009.001560

42. Silverman, G. M., Gosho, C., Tran, T., Kahn, M. G., & Haendel, M. A. (2024). Developing a natural language processing system using transformer-based models for adverse drug event detection in electronic health

records. *medRxiv preprint*. https://doi.org/10.1101/2024.07.09.24310100

43. Styler IV, W. F., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P. C., Erickson, B., Miller, T., Lin, C., Savova, G., & Pustejovsky, J. (2014). Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, *2*, 143–154. https://doi.org/10.1162/tacl_a_00172

44. Sun, W., Rumshisky, A., & Uzuner, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, *20*(5), 806–813. https://doi.org/10.1136/amiajnl-2013-001628

45. Uzuner, O., Goldstein, I., Luo, Y., & Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, *15*(1), 14–24. https://doi.org/10.1197/jamia.M2408

46. Uzuner, O., Stubbs, A., & Filannino, M. (2018). *n2c2 2018 — Track 2: Adverse drug events and medication extraction in EHRs*. https://portal.dbmi.hms.harvard.edu/projects/n2c2-2018-t2/

47. van Diepen, M., Ramspek, C. L., Jager, K. J., Zoccali, C., & Dekker, F. W. (2024). Detection of adverse drug events in Dutch clinical free text documents using transformer models: Benchmark study. *arXiv preprint*. https://arxiv.org/abs/2507.19396

48. Wang, J., Abu-el-rub, N., Gray, J., Pham, H. A., Zhou, Y., Manion, F. J., Liu, M., Song, X., Xu, H., & Rouhizadeh, M. (2024). Evaluating the performance of large language models for named entity recognition in ophthalmology clinical free-text notes. *Ophthalmology Science*, *4*(6), 100536. https://doi.org/10.1016/j.xops.2024.100536

49. Wei, Q., Ji, Z., Li, Z., Du, J., Wang, J., Xu, J., Xiang, Y., Tiryaki, F., Wu, S., Zhang, Y., Tao, C., & Xu, H. (2020). A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *Journal of the American Medical Informatics Association*, *27*(1), 13–21. https://doi.org/10.1093/jamia/ocz063

50. Wu, S., Miller, T., Masanz, J., Coarr, M., Halgrim, S., Carrell, D., & Clark, C. (2014). Negation's not solved: Generalizability versus optimizability in clinical natural language processing. *PLOS ONE*, *9*(11), e112774. https://doi.org/10.1371/journal.pone.0112774

51.