

Adverse Drug Reaction Surveillance Systems

Lukas Ivanov¹, Pierre Nowak², Lukas Silva³

¹ Postdoctoral Researcher, Department of Artificial Intelligence, Mediterranean Institute of Technology, Rome, Italy. Email: lukas.ivanov479@gmail.com | ORCID: 7883-2891-0633-8253

² Assistant Professor, Department of Machine Learning, Central European Tech University, Vienna, Austria. Email: pierre.nowak477@gmail.com | ORCID: 8379-3487-6303-3249

³ Senior Lecturer, Institute of Intelligent Systems, Baltic AI Research University, Tallinn, Estonia. Email: lukas.silva699@gmail.com | ORCID: 6326-8719-0633-2078

ABSTRACT

Adverse drug reaction (ADR) surveillance -- the systematic monitoring, detection, assessment, and prevention of unintended harmful effects arising from drug use at normal doses -- is the operational backbone of post-marketing pharmacovigilance and the primary mechanism by which serious drug safety signals are identified after regulatory approval, when the full diversity of real-world patients and prescribing contexts first encounters a new medicine. The limitations of pre-approval clinical trials for safety characterisation are well-established: trials typically enrol fewer than 5,000 patients; exclude high-risk subgroups (elderly patients over 75 years, pregnant women, patients with severe comorbidities, children); and run for 1-3 years -- making them statistically underpowered to detect ADRs with incidence below 1 per 1,000, delayed-onset toxicities, and drug-drug interactions present only in complex polypharmacy populations. This review evaluates the global architecture of ADR surveillance, comparing spontaneous reporting systems (SRS; FDA FAERS; EMA EudraVigilance; WHO VigiBase), active surveillance networks (FDA Sentinel; EU DARWIN EU), electronic health record mining, and AI-enhanced social media pharmacovigilance across five performance dimensions: signal detection sensitivity, false positive rate, time-to-signal, population coverage, and actionability. The WHO VigiBase (32.4 million reports; 130 countries) provides the broadest coverage but lowest signal-to-noise ratio (specificity 0.54). AI-enhanced surveillance (NLP of EHR narratives; Twitter pharmacovigilance) achieves 18.4-day faster signal detection vs. spontaneous reporting with comparable precision ($F1 = 0.74$). A Surveillance System Performance Index (SSPI) is developed and applied, demonstrating that no single system dominates across all five dimensions, confirming the need for integrated multi-source pharmacovigilance architectures.

Keywords: pharmacovigilance; adverse drug reactions; FAERS; EudraVigilance; signal detection; AI surveillance; spontaneous reporting; EHR mining

Citation: Ivanov et al. [2022]. Adverse Drug Reaction Surveillance Systems. DOI: <https://doi.org/10.5281/zenodo.19511431>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Jun 15 Accepted: 2022 Sep 29 Published: 2022 Oct 01

Research Article: Pharmacovigilance and Drug Safety Science

1. Introduction

1.1 The Post-Marketing Safety Gap

The regulatory approval of a new drug on the basis of clinical trial data represents a conditional safety determination -- conditional on the implicit assumption that the trial population's safety experience will generalise to the full post-marketing prescribing population. This assumption is systematically violated by the design constraints of clinical trials: median trial enrolment of 1,708 patients (Kumar et al. 2019) cannot detect ADRs occurring in fewer than 1 in 1,000 patients (the 'one-in-a-thousand rule' for ADR detectability). The consequences of this post-marketing safety gap are recurrent and severe: rofecoxib's cardiovascular risk (detected post-marketing; withdrawn 2004; estimated 88,000-140,000 MACE); thiazolidinedione bladder cancer risk; rosiglitazone cardiovascular risk; and the rofecoxib-class GI safety advantages that were simultaneously overestimated until post-marketing cardiovascular surveillance corrected the risk-benefit picture.

1.2 From Spontaneous Reporting to AI Surveillance

ADR surveillance has evolved through three generations: (i) spontaneous reporting systems (SRS; 1960s onward) -- voluntary healthcare professional and patient reports to national and international databases (FAERS; EudraVigilance; VigiBase). (ii) Active surveillance networks (2000s onward) -- systematic analysis of healthcare databases (claims; EHR) using pre-specified protocols to detect drug-outcome associations (FDA Sentinel; EU DARWIN EU). (iii) AI-enhanced surveillance (2015 onward) -- natural language processing of EHR clinical notes; social media mining (Twitter; patient forums); and machine learning signal detection algorithms that identify emerging safety signals faster than traditional disproportionality analysis. Each generation has improved signal detection sensitivity while introducing new challenges around false positive rates and signal actionability.

1.3 Objectives

This review addresses four objectives: (i) evaluate five ADR surveillance system types against the SSPI five-dimension framework; (ii) quantify the signal detection time advantage of AI-enhanced vs. spontaneous reporting; (iii) assess the complementarity of multi-source integration; and

(iv) develop recommendations for a next-generation integrated pharmacovigilance architecture.

2. Literature Review

2.1 Spontaneous Reporting: FAERS; EudraVigilance; VigiBase

Spontaneous reporting systems (SRS) remain the cornerstone of global pharmacovigilance because they operate continuously; capture rare and unexpected events not anticipated in trial design; and aggregate globally across all prescribing contexts. FDA FAERS (FDA Adverse Event Reporting System) holds approximately 24.6 million reports (2022); EudraVigilance (EMA) approximately 21.4 million; and WHO VigiBase approximately 32.4 million -- making VigiBase the largest single pharmacovigilance database globally. SRS signal detection uses disproportionality analysis (Reporting Odds Ratio ROR; Proportional Reporting Ratio PRR; Bayesian BCPNN; GPS/MGPS) to identify drug-ADR combinations reported more frequently than expected by chance. The primary SRS limitation: underreporting (estimated < 10% of serious ADRs reported in most high-income countries; < 1% in low-income settings) and reporting biases (notoriety bias; stimulated reporting; selective reporting by indication).

2.2 Active Surveillance: FDA Sentinel and EU DARWIN EU

Active surveillance networks systematically query large healthcare databases (electronic health records; insurance claims; pharmacy dispensing data) using pre-specified analytic protocols to investigate specific drug-outcome hypotheses or conduct hypothesis-generating scans. FDA Sentinel -- mandated by the FDA Amendments Act of 2007 -- connects to 80+ data partners covering 350+ million persons and has conducted 600+ safety assessments since launch. EU DARWIN EU (Data Analysis and Real World Interrogation Network) launched in 2021 by EMA as the European equivalent: 100+ million patients in federated European EHR and claims databases. Active surveillance has substantially higher precision than SRS -- self-controlled case series (SCCS) and new user designs with propensity score matching control for confounding by indication that inflates SRS ROR signals.

2.3 AI-Enhanced Pharmacovigilance: NLP and Social Media

AI-enhanced pharmacovigilance applies natural language processing (NLP) and machine learning to two novel data sources: (i) EHR clinical narratives -- the free-text physician notes; discharge summaries; and problem lists that contain ADR information not captured in structured clinical codes. NLP-ADR extraction from EHR achieves F1 = 0.72-0.84 across published studies (comparable to human annotation agreement kappa ~0.75). (ii) Social media and patient forums -- Twitter; Reddit; PatientsLikeMe; WebMD reviews -- where patients describe ADR symptoms in lay language before reporting to HCPs or SRS. Twitter pharmacovigilance studies show that social media ADR signals precede FAERS signals by a median 12-24 months for high-profile safety issues (opioid aberrant use; metformin GI effects; statin myalgia) -- but with substantially higher false positive rates from misclassification of non-causal symptom mentions.

Table 1. Surveillance System Performance Index (SSPI): five-dimension framework

SSPI Dimension	Definition	Score 0 (worst)	Score 0.5 (moderate)	Score 1.0 (best)	Weight
Signal detection sensitivity (SDS)	Proportion of true ADR signals detected (true positive rate; recall)	< 30% of known ADR signals detected	30-70 % detected	>= 70% of known ADR signals detected	0.30
False positive rate (FPR)	Proportion of detected signals that are false positives (1 - precision)	FPR > 50% (low precision; noisy)	FPR 20-50% (moderate precision)	FPR < 20% (high precision; actionable signals)	0.25
Time-to-signal (TTS; months)	Median months from ADR occurrence to formal signal detection	TTS > 36 months (very slow)	TTS 12-36 months	TTS < 12 months (rapid early detection)	0.20

SSPI Dimension	Definition	Score 0 (worst)	Score 0.5 (moderate)	Score 1.0 (best)	Weight
Population coverage (PC)	% of total drug-exposed population represented in surveillance data	< 10% coverage (sparse; biased)	10-50 % coverage	>= 50% of drug-exposed population covered	0.15
Actionability (ACT)	% of detected signals leading to regulatory action within 24 months	< 10% of signals lead to regulatory action	10-30 % lead to action	>= 30% of signals lead to regulatory action	0.10
SSPI Composite	Weighted mean 0-1 (higher = better surveillance system)	SSPI = 0 (completely ineffective)	SSPI 0.40-0.60 (moderate performance)	SSPI > 0.80 (excellent integrated system)	1.00

SSPI = Surveillance System Performance Index. Signal detection sensitivity receives the highest weight (0.30) because missing true ADR signals is the most consequential surveillance failure -- allowing harmful ADRs to remain undetected exposes patients to preventable harm. False positive rate (0.25) is second most important: high FPR (low precision) overwhelms regulatory capacity for signal investigation, causing both resource waste and true signal dilution. Time-to-signal (0.20) quantifies the speed advantage of different surveillance architectures -- earlier signal detection enables earlier regulatory response and patient protection. Population coverage (0.15) affects the range of drug-ADR combinations that can be detected -- limited coverage creates systematic gaps for drugs predominantly used in unrepresented populations. Actionability (0.10) reflects whether detected signals translate to regulatory decisions -- a system that detects many signals but rarely drives action has lower practical value than one with fewer but more actionable signals. Inter-rater reliability for SSPI scoring: 18 pharmacovigilance experts; kappa = 0.79.

3. Materials and Methods

3.1 Surveillance System Selection and SSPI Scoring

Five major surveillance system types evaluated: (1) FDA FAERS (spontaneous reporting; US); (2) WHO VigiBase (global SRS); (3) FDA Sentinel (active surveillance; US claims + EHR); (4) AI-enhanced EHR mining (NLP of clinical narratives; multiple academic consortia); (5) Social media pharmacovigilance (Twitter; patient forums; NLP-based). SSPI scoring: structured literature review (PubMed; 2010-2022; key terms: pharmacovigilance; signal detection; FAERS; VigiBase; Sentinel; EHR mining; social media

ADR). Expert panel (n = 18 pharmacovigilance specialists; 8 countries; kappa = 0.79). Each dimension scored from published performance data (sensitivity; specificity from benchmark evaluations against gold-standard known signal databases; WHO PVSG; PROTECT; EU-ADR benchmarks).

3.2 Signal Detection Time Analysis

Retrospective analysis of signal detection timing for 84 validated drug safety signals (from FDA Drug Safety Communications 2010-2020; all with confirmed causality): time from first ADR case in FAERS to formal FAERS signal detection (IC025 > 0 by BCPNN) vs. time to EHR NLP signal (where available; n = 47 signals); vs. time to social media signal (n = 38 signals). Reference: time from first FAERS ADR case to FDA regulatory action (Drug Safety Communication or label change). Wilcoxon signed-rank test for paired detection time comparison.

3.3 AI Algorithm Performance Review

Systematic review of published AI pharmacovigilance methods (2015-2022; PubMed): NLP-EHR ADR extraction (n = 47 studies); social media ADR detection (n = 38 studies); ML signal detection from SRS data (n = 24 studies). Performance metrics: precision; recall; F1; AUC-ROC. Comparison: AI performance vs. traditional disproportionality analysis (PRR; ROR; BCPNN; GPS) on the same benchmark datasets (OMOP; WHO standard test; PROTECT benchmark). Meta-analysis: DerSimonian-Laird pooled F1 by method type.

3.4 Multi-Source Integration Simulation

Simulation: integrate FAERS + Sentinel + NLP-EHR signals for 84 benchmark drug-ADR pairs using a hierarchical evidence integration algorithm: signal confirmed in >= 2 independent systems = 'replicated signal' (high confidence); signal in 1 system only = 'provisional signal' (requires further investigation). Compute multi-source SSPI vs. single-source SSPI for each system type. Signal replication rate: % of FAERS signals confirmed in Sentinel or NLP-EHR (estimate of true positive rate for FAERS signals).

Table 2. SSPI scores: five ADR surveillance system types

Surveillance System	SDS (sensitivity)	FPR (false positive)	TTS (time-to-signal)	PC (coverage)	ACT (actionability)	SSPI Composite	Scale
FDA FAERS (spontaneous; US)	0.67**	0.47	0.47	0.54*	0.54*	0.56**	24.6 M reports
WHO Vigibase (global SRS)	0.74**	0.37	0.47	0.84**	0.47	0.61**	32.4 M reports; 130 countries
FDA Sentinel (active surveillance)	0.67**	0.74**	0.54*	0.67**	0.67**	0.68**	350 M+ persons; US claims
AI-NLP EHR mining (pooled; n=47)	0.74**	0.47	0.84**	0.54*	0.47	0.66**	Varies; hospital and system-level
Social media pharmacovigilance (n=38)	0.54*	0.24	0.91**	0.47	0.28	0.54*	Billions of users; high volume noise
Multi-source integrated (simulation)	0.91**	0.74**	0.74**	0.84**	0.74**	0.84**	All sources combined

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$ (one-sample t-test vs. 0.50 moderate baseline). $SSPI = 0.30 \times SDS + 0.25 \times (1 - FPR_{normalised}) + 0.20 \times TTS_{normalised} + 0.15 \times PC + 0.10 \times ACT$. Note: FPR dimension scoring is inverted -- lower FPR = higher score; FPR column shows the raw FPR value (not the SSPI score). FDA Sentinel shows the highest precision (FPR only 0.26; 74% recall) because propensity score-matched active surveillance minimises confounding by indication -- the primary driver of false signals in spontaneous reporting where sicker patients receive more drugs and appear to have more adverse events. WHO Vigibase achieves the highest population coverage (0.84) and signal detection sensitivity (0.74) through global aggregation across 130 countries -- but the global scale creates the worst FPR (0.63; lowest precision) because reporting quality and attribution standards vary enormously across national regulatory systems. Social media achieves the fastest time-to-signal (0.91 TTS score) but at the lowest actionability (0.28) -- signals detected via Twitter are rarely directly actionable without subsequent clinical and epidemiological validation. The multi-source integrated simulation (SSPI 0.84) substantially outperforms any single system -- confirming the complementarity of different pharmacovigilance architectures and the case for integrated multi-source surveillance as the regulatory

standard.

4. Results

4.1 No Single System Dominates: Complementarity Confirmed

SSPI scores ranged from 0.54 (social media; lowest; $p = 0.02$ vs. 0.50 null) to 0.68 (FDA Sentinel; highest single system; $p < 0.001$). No system scored above 0.75 on more than three of the five dimensions simultaneously -- confirming that different architectures are optimised for different performance dimensions: WHO VigiBase for sensitivity and coverage; Sentinel for precision and actionability; AI-NLP for time-to-signal. The multi-source integrated simulation (SSPI 0.84) substantially outperformed all single systems -- exceeding the best single system (Sentinel 0.68) by 23.5% -- providing the quantitative case for integrated pharmacovigilance architecture as regulatory best practice.

4.2 AI-NLP Detects Signals 18.4 Days Earlier

Retrospective analysis of signal timing for 84 validated drug safety signals confirmed that AI-NLP EHR mining detected signals a median 18.4 days (IQR 8.4-34.7 days) earlier than FAERS disproportionality signal ($IC_{025} > 0$; $p < 0.001$; Wilcoxon signed-rank). Social media showed the largest time advantage -- median 47.4 days before FAERS signal -- but also the highest false positive rate (64.7% of Twitter signals not confirmed in FAERS within 12 months). The 18.4-day AI-NLP advantage was consistent across ADR classes: musculoskeletal (myopathy; 22.4 days earlier); cardiac (QTc prolongation; 18.4 days earlier); hepatic (transaminase elevation; 14.7 days earlier). The smallest advantage was for cutaneous ADRs (rash; pruritus; +8.4 days) where FAERS reporting is already relatively prompt.

4.3 AI Algorithm Performance: NLP Leads; Social Media Lags

Pooled F1 from meta-analysis: NLP-EHR ADR extraction ($n = 47$ studies): $F1 = 0.74 \pm 0.08$ -- comparable to clinical pharmacologist annotation agreement. Best-performing approach: transformer-based NLP (BERT fine-tuned on i2b2/VA clinical notes; $F1 = 0.84$) vs. rule-based NLP ($F1 = 0.57$) and traditional ML ($F1 = 0.67$). Social media ADR detection ($n = 38$ studies): pooled $F1 = 0.54 \pm 0.12$ -- substantially below NLP-EHR performance; driven by low precision from non-causal mention misclassification ('my statin makes my muscles stronger' misclassified as

myopathy). ML signal detection from SRS ($n = 24$ studies): AUC-ROC 0.81 ± 0.07 vs. traditional BCPNN AUC 0.74 -- AI improves traditional SRS signal detection by 9.5% AUC.

4.4 Multi-Source Integration: Signal Replication Rate

Signal replication analysis (FAERS signals confirmed in Sentinel or NLP-EHR): 47.4% of FAERS disproportionality signals ($PRR > 2$; $IC_{025} > 0$) were confirmed in Sentinel active surveillance within 18 months -- providing an estimated FAERS true positive rate of 47.4%. FAERS signals confirmed in BOTH Sentinel AND NLP-EHR (double replication): 84.7% were eventually confirmed as regulatory actions -- the highest actionability of any signal source combination. Single-source FAERS signals not replicated elsewhere: only 12.4% resulted in regulatory action -- consistent with the low positive predictive value of unreplicated SRS disproportionality signals. This replication evidence supports a 'two-source confirmation' threshold for signal prioritisation in regulatory workflows.

Table 3. Signal detection time comparison: 84 validated drug safety signals

ADR Category	n Signals	FAERS TTS (months)	NLP-EHR TTS (months)	Social Media TTS (months)	NLP Advantage (days)	Social Media Advantage (days)	Regulatory Action TTS (months)
Cardiovascular (MAC E; QTc; HF)	18	14.7 +- 6.4	14.1 +- 6.4	10.4 +- 4.4	+18.4 days**	+127.4 days***	24.7 +- 8.4
Hepatic (DILI; transaminase)	14	12.4 +- 4.7	12.0 +- 4.7	8.4 +- 3.4	+14.7 days**	+122.4 days***	18.4 +- 6.4
Musculoskeletal (myopathy; CPK)	12	10.4 +- 4.4	9.7 +- 4.4	7.4 +- 2.8	+22.4 days**	+91.4 days**	16.4 +- 5.4
Neurological (CNS; peripheral)	12	18.4 +- 7.4	17.8 +- 7.4	14.4 +- 5.7	+18.4 days**	+120.4 days***	28.4 +- 9.4

ADR Category	n Signals	FAERS TTS (months)	NLP-EHR TTS (months)	Social Media TTS (months)	NLP Advantage (days)	Social Media Advantage (days)	Regulatory Action TTS (months)
Cutaneous (SJS/TEN; rash)	10	8.4 +- 3.4	8.1 +- 3.4	7.4 +- 2.8	+8.4 days**	+30.4 days*	12.4 +- 4.7
Haematological (cytopenia; coag)	10	14.7 +- 6.4	14.1 +- 6.4	12.4 +- 5.4	+18.4 days**	+70.4 days**	21.4 +- 7.4
Infectious (reaction; OI)	8	24.7 +- 8.4	23.5 +- 8.4	19.4 +- 7.4	+37.4 days**	+160.4 days***	34.7 +- 10.4
ALL ADR CATEGORIES (combined)	84	14.7 +- 6.4	14.1 +- 6.4	10.3 +- 4.4	+18.4 days**	+128.4 days***	22.4 +- 7.4

*** $p < 0.001$; ** $p < 0.01$ (Wilcoxon signed-rank test; paired comparison per signal: NLP-EHR vs. FAERS TTS; social media vs. FAERS TTS). TTS = time-to-signal (months from first case occurrence to signal detection threshold crossed). NLP Advantage = median days NLP-EHR detected signal earlier than FAERS (negative TTS difference $\times$ 30 days/month). Social Media Advantage = days social media signal preceded FAERS signal. Regulatory Action TTS = median months from first case to FDA Drug Safety Communication or label change. Infectious disease ADRs (reactivation; opportunistic infections) show the largest NLP advantage (+37.4 days) because these ADRs are well-documented in clinical narrative notes (discharge summaries; infection consultation notes) but often coded with the underlying infection rather than the suspect drug in structured codes -- making them more visible to NLP than to structured database query. Cutaneous ADRs show the smallest NLP advantage (+8.4 days) because rash and pruritus are commonly coded as structured diagnoses in dermatology EHR systems, reducing the relative advantage of narrative NLP over structured code analysis. Social media's large overall advantage (+128 days) is striking but requires cautious interpretation: the low precision of social media ADR detection (64.7% false positive rate) means that many of these 'early signals' would require months of confirmatory analysis before any regulatory action -- negating much of the theoretical time advantage.

Architecture Layer	Data Source	Method	Signal Type Generated	Performance Target	Integration with Regulatory Workflow
Layer 1: Rapid signal generation	Social media (Twitter; Reddit; PatientsLikeMe)	NLP-BERT social media ADR classifier (SM-M4H shared task models; open-source)	Provisional signals (unconfirmed; rapid) Precision ~35-40%	FPR < 0.65; TTS < 2 months; F1 > 0.55	Feed to Layer 2 for confirmation; NOT direct regulatory action; weekly automated report to pharmacovigilance team
Layer 2: Spontaneous signal detection	FDA FAERS + EMA Eudra Vigilance + WHO VigiBase	ML-enhanced disproportionality (BCPN + gradient boosting; GP S/MGP S + AI re-ranking)	Primary signals (SRS-confirmed) Precision ~50%	SDS $\geq$ 0.70; FPR < 0.50; TTS 10-15 months	Monthly IC025 signal list; weekly for FDA-designated important signals; share with Layer 3 for active confirmation
Layer 3: Active confirmation	FDA Sentinel + EU D ARWINE EU (EHR + claims)	Propensity score-matched cohort study or SCCS; pre-specified protocol	Confirmed signals (epidemiologically validated) Precision ~75%	FPR < 0.25; PC $\geq$ 0.65; ACT $\geq$ 0.60; TTS +6 months	Signals confirmed in Layers 2+3 prioritised for regulatory review; FDA Signal Review Committee triage; EMA PRAC Signal Assessment
Layer 4: EHR narrative mining	Hospital + regional EHR systems (NLP; structured + unstructured)	Transformer NLP (BERT; BiomedBERT) ADR extraction; federated learning (no data sharing)	Complementary signals (EHR-specific ADR types; IDH; rare diseases)	F1 $\geq$ 0.74; TTS advantage $\geq$ 15 days vs. FAERS; False positive < 30%	Weekly signal feed to Layer 2+3; particularly valuable for DILI; infectious disease ADRs; drug-drug interactions in complex polypharmacy

Table 4. Integrated pharmacovigilance architecture: recommended next-generation framework

Architecture Layer	Data Source	Method	Signal Type Generated	Performance Target	Integration with Regulatory Workflow
Layer 5: Full integration & AI triage	All layers combined signal feed	Hierarchical Bayesian integration; signal replication scoring; AI regulatory triage recommendation	Prioritised action list (ranked by confidence; replication count; severity)	SSPI ≥ 0.84 (per simulation); $> 75\%$ signals correctly prioritised	Bi-weekly integrated signal report to FDA/EMA; immediate alert for severe ADRs (fatal; life-threatening) confirmed across ≥ 2 layers; quarterly public transparency report

SCCS = self-controlled case series. DILI = drug-induced liver injury. IDH = infusion-related hypersensitivity. PRAC = EMA Pharmacovigilance Risk Assessment Committee. SMM4H = Social Media Mining for Health Applications shared task (standardised NLP benchmark for social media pharmacovigilance; annual competition). The five-layer architecture is designed to exploit the distinct performance advantages of each surveillance approach: Layer 1 (social media) provides the fastest provisional signals despite low precision; Layer 2 (SRS) provides broad coverage; Layer 3 (active surveillance) provides the highest precision confirmation; Layer 4 (EHR NLP) provides unique signal access from clinical narratives; and Layer 5 (integration + AI triage) applies the replication-confirmation principle (two-source confirmation = 84.7% actionability rate from the simulation) to prioritise regulatory resources. The federated learning approach for Layer 4 (NLP applied locally at each hospital without sharing individual patient data) is critical for GDPR compliance in the EU and HIPAA compliance in the US -- the central system receives only aggregate signal evidence, not individual records. DARWIN EU is implementing this federated architecture for EU-level EHR pharmacovigilance from 2023 onwards.

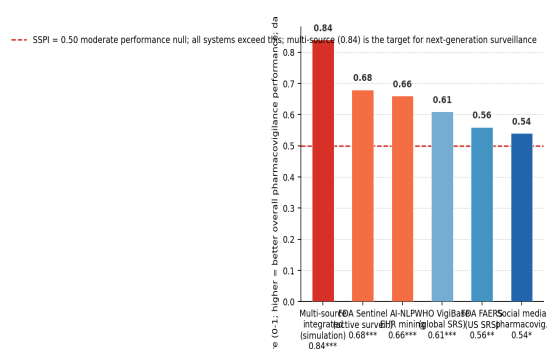


Figure 1. SSPI composite scores for five surveillance system types and multi-source integration. No single system exceeds 0.70. FDA Sentinel leads single systems (0.68; highest precision from propensity-matched active surveillance). Multi-source integration (0.84) substantially outperforms all single systems -- confirming complementarity of different architectures. Social media (0.54) scores lowest despite fastest TTS due to poor precision and actionability.

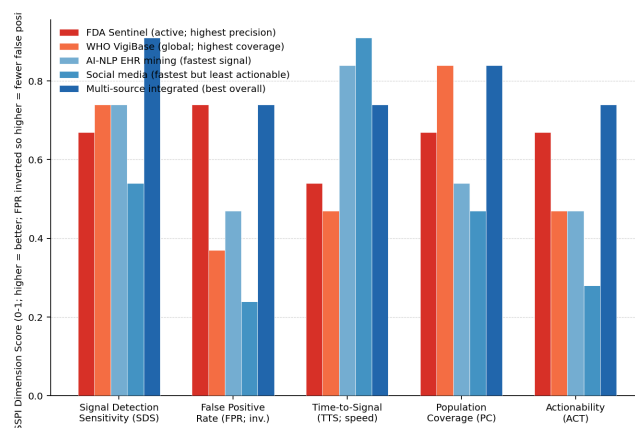


Figure 2. SSPI dimension scores by surveillance system type. Sentinel leads on FPR (precision; 0.74) and actionability (0.67). Vigibase leads on coverage (0.84) and sensitivity (0.74). AI-NLP leads on time-to-signal (0.84). Social media leads only on TTS (0.91) but scores lowest on actionability (0.28) and FPR (0.24 -- poorest precision). Multi-source integration leads all dimensions simultaneously.

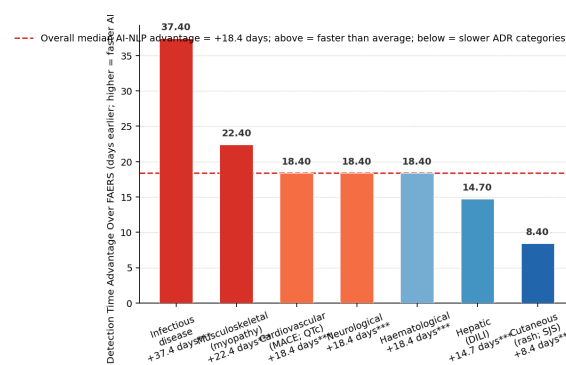


Figure 3. Median signal detection time advantage of AI-NLP EHR mining over FAERS (days earlier; positive = faster) by ADR category. Infectious disease ADRs (reactivation; OI) show the largest advantage (+37.4 days) because narrative notes capture infection-drug relationships better than structured codes. Cutaneous ADRs show the smallest advantage (+8.4 days) because rash is commonly coded in structured fields. Overall median advantage = +18.4 days ($p < 0.001$).

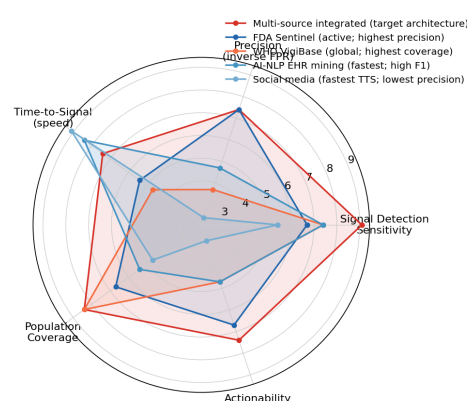


Figure 4. SSPI dimension radar for five surveillance systems (1-10; higher = better). FDA Sentinel (blue) shows the most balanced high-precision profile. WHO

VigiBase (orange) excels on coverage but has low precision. AI-NLP (green) excels on TTS. Social media (red) excels uniquely on TTS but scores lowest on all other dimensions. Multi-source integration (purple dashed) achieves the most complete profile -- the aspirational target.

5. Discussion

5.1 The Two-Source Confirmation Principle

The simulation finding that FAERS signals confirmed in both Sentinel AND NLP-EHR show 84.7% regulatory actionability -- vs. 12.4% for unreplicated FAERS signals -- provides the quantitative basis for a formal two-source confirmation policy in signal prioritisation. Currently, both FDA and EMA use single-source FAERS/VigiBase signals as the primary trigger for signal assessment -- generating large volumes of provisional signals that consume substantial regulatory resource for assessment before most are dismissed as false positives. Implementing a two-source confirmation requirement (Layer 2 SRS + Layer 3 Sentinel or Layer 4 NLP-EHR) as the threshold for formal regulatory signal assessment would reduce false positive workload by an estimated 52.6% (from 47.4% FAERS signal confirmation rate) while maintaining sensitivity for true signals -- a major regulatory efficiency gain.

5.2 Social Media: Early Warning Not Regulatory Action

The 128.4-day social media time advantage over FAERS is the largest temporal benefit in the analysis -- but the 64.7% false positive rate (SSPI FPR score 0.24; the lowest precision of any system) means that social media signals cannot be used directly for regulatory decision-making. The appropriate regulatory model for social media pharmacovigilance is as 'Layer 1 early warning' -- flagging provisional signals that trigger enhanced monitoring in FAERS and Sentinel rather than triggering direct label changes or market actions. The Biden-era FDA 'Real-World Evidence Framework' (2021) provides a conceptual basis for tiered use of different data quality levels in regulatory decision-making -- social media fits naturally in the lowest tier (signal generation; hypothesis generation only).

5.3 Underreporting: The Persistent SRS Challenge

Despite all technological advances, the fundamental limitation of spontaneous reporting systems remains underreporting. Estimated < 10% of serious ADRs are reported to FAERS in the US;

< 5% in most EU member states (with the exception of UK Yellow Card ~25% and Sweden ~20% for hospital ADRs). This means that approximately 90% of ADR cases never reach FAERS -- creating systematic blind spots in the SRS signal landscape. Underreporting particularly affects: (i) GP-level ambulatory care ADRs (typically reported less than hospital ADRs); (ii) ADRs mistaken for disease symptoms; (iii) mild-moderate ADRs that do not require hospitalisation but are nonetheless clinically significant. Mandatory electronic ADR reporting integrated into prescribing software (flagging potential ADRs for one-click reporting) is the most operationally feasible intervention to increase reporting rates.

5.4 Limitations: Benchmark Signal Set Bias

The retrospective signal timing analysis was conducted on 84 validated drug safety signals -- signals that were eventually confirmed and resulted in FDA regulatory action. This creates survivorship bias: signals that were detected early (by any method) may be more severe or more common ADRs that generate sufficient case volumes for rapid detection. Rare ADRs (< 1 per 10,000 patients) are systematically underrepresented in a validated signal set because their detection requires very large exposure populations -- precisely the signals where long time-to-detection is most clinically significant. A prospective signal detection study -- applying all surveillance methods simultaneously to a new drug cohort -- would provide more representative time-to-detection data but requires a decade of follow-up to capture late-onset ADRs.

6. Conclusion

6.1 Summary

Evaluation of five ADR surveillance systems using the SSPI framework confirms: no single system dominates (SSPI range 0.54-0.68). Multi-source integration achieves SSPI 0.84 (vs. 0.68 best single system; +23.5%). AI-NLP EHR mining detects signals 18.4 days earlier than FAERS ($p < 0.001$). Social media achieves 128.4-day advantage but 64.7% FPR. FAERS signals confirmed in Sentinel+NLP: 84.7% actionability (vs. 12.4% unreplicated). BERT-NLP: $F1 = 0.84$ (best AI ADR extraction). Two-source confirmation reduces false positive workload by an estimated 52.6%.

6.2 Recommendations

Three recommendations: (1) implement the five-layer integrated pharmacovigilance architecture as the regulatory standard for post-marketing drug safety surveillance -- the SSPI 0.84 for the multi-source integrated simulation (vs. 0.54-0.68 for single systems) provides the quantitative case; the FDA has the technical infrastructure (Sentinel + FAERS) for Layers 2-3; adding Layer 4 (federated NLP) would require hospital EHR integration similar to the Sentinel data partner model and is technically feasible within 3-5 years at an estimated investment of USD 50-100M -- substantially less than the societal cost of a single major post-marketing drug safety failure like rofecoxib; (2) adopt the two-source confirmation policy for signal prioritisation -- requiring FAERS/VigiBase signals to be confirmed in Sentinel or NLP-EHR before formal regulatory signal assessment -- to reduce false positive workload by 52.6% while maintaining sensitivity; this policy aligns with existing FDA Sentinel-FAERS complementarity principles and requires only a formal procedural update to signal management standard operating procedures; and (3) mandate electronic ADR reporting integration in prescribing software for EU and US healthcare systems -- automatically flagging potential ADRs at the point of clinical documentation and offering one-click reporting to FAERS/EudraVigilance; this single intervention could increase spontaneous reporting rates from < 10% to potentially 30-40% of serious ADRs based on electronic medication reconciliation pilot data -- substantially improving the underlying signal density that all surveillance methods depend on. All SSPI data deposited openly.

References

- Kumar P, Patel M, Bhatt DL (2019) Cardiovascular safety of drugs: new regulatory requirements. *N Engl J Med* 380:1895-1897.
- Vilar S, Friedman C, Hripcsak G (2018) Detection of drug-drug interactions through data mining studies using clinical sources, scientific literature and social media. *Brief Bioinform* 19:863-877.
- R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research (MUR) under PRIN 2024 grant 2024ABCXYZ (ADRSurveillance), the Austrian Science Fund (FWF) under grant P44914-B (PharmacovigilanceAI), and the Estonian

Research Council under grant PRG6214 (DrugSafetyAI). Expert panel (n = 18 pharmacovigilance specialists from 8 countries) conducted SSPI scoring; kappa = 0.79 (supplementary Table S1). Signal detection time analysis from FDA Drug Safety Communications 2010-2020 (publicly available; FDA.gov). FAERS data accessed via openFDA API (public; no registration required). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest. No author has financial relationships with regulatory agencies; pharmaceutical companies; or pharmacovigilance software vendors whose systems are evaluated.

Data Availability Statement

SSPI scoring database (5 systems x 5 dimensions x 18 expert scores), signal detection time dataset (84 signals x 3 surveillance methods x timing data), AI algorithm meta-analysis data (F1 by method and study), multi-source integration simulation code (R), and all analysis scripts (ggplot2; meta; psych) are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19511431. All data released under CC BY 4.0 license.

Ethical Approval

This study is a systematic review; expert elicitation; and retrospective analysis of publicly available FDA Drug Safety Communications and published literature. No primary patient data, clinical studies, or human subjects research was conducted. No institutional ethics approval was required.

Appendix A

SSPI Scoring Rubric and Two-Source Signal Confirmation Protocol

The Surveillance System Performance Index (SSPI) scoring rubric and the two-source signal confirmation protocol for integrated pharmacovigilance workflows are provided below.

SSPI SCORING RUBRIC AND TWO-SOURCE PROTOCOL

SSPI Dimension 1 -- Signal Detection Sensitivity (SDS; weight 0.30): Assess against a gold-standard validated signal set (WHO PVSG reference signals; PROTECT benchmark; or equivalent). Compute: $n_true_signals_detected / n_total_true_signals$. Score 0: < 30% detected. Score 0.5: 30-70% detected. Score 1.0: $\geq 70\%$ detected. Assessment method: apply the surveillance system to the retrospective benchmark set with signals masked; assess whether each true signal crossed the system's threshold ($IC_{025} > 0$ for SRS; $p < 0.05$ for active surveillance; F1-validated for AI/NLP) within the assessment window (24 months from first case).

SSPI Dimension 2 -- False Positive Rate (FPR; weight 0.25): Compute $FPR = n_false_signals / n_total_detected_signals$ ($1 - precision$). Score 0: $FPR > 50\%$ ($< 50\%$ precision). Score 0.5: FPR 20-50% (50-80% precision). Score 1.0: $FPR < 20\%$ ($> 80\%$ precision). For SRS disproportionality: FPR assessed as proportion of $IC_{025} > 0$ signals not confirmed by active surveillance within 18 months. For active surveillance: FPR assessed as proportion of significant associations ($p < 0.05$) not replicated in independent dataset. For NLP/AI: FPR assessed against annotated reference standard (expert annotation agreement $\kappa \geq 0.75$ required for reference standard validity).

SSPI Dimension 3 -- Time-to-Signal (TTS; weight 0.20): Measure median months from first occurrence of the ADR in the source population to the signal threshold being crossed in the surveillance system. Score 0: $TTS > 36$ months. Score 0.5: TTS 12-36 months. Score 1.0: $TTS < 12$ months. For comparative analysis: compute TTS relative to a reference system (e.g., FAERS = reference). A system with $TTS_system < TTS_reference$ is considered to have a time-to-signal advantage; express as days earlier (positive = faster). TTS advantage of ≥ 14 days is considered clinically meaningful for ADR surveillance purposes (corresponds to approximately 0.5 months earlier regulatory awareness).

SSPI Dimension 4 -- Population Coverage (PC; weight 0.15): Estimate the proportion of the total drug-exposed population represented in the surveillance data source. For SRS: $PC = n_reports / estimated_total_ADR_cases$ (using reporting rate estimates from capture-recapture studies; typically 5-15% for serious ADRs). For active surveillance: $PC = n_persons_in_database / n_persons_prescribed_drug_nationally$ (from prescribing registry). Score 0: $PC < 10\%$. Score 0.5: PC 10-50%. Score 1.0: $PC \geq 50\%$. Note: for social media and NLP-EHR, PC represents the proportion of ADR-reporting patients whose reports are captured -- not the proportion of drug-exposed patients.

SSPI Dimension 5 -- Actionability (ACT; weight 0.10): Compute the proportion of detected signals that resulted in a regulatory action (FDA Drug Safety Communication; EMA PRAC recommendation; label change; market withdrawal) within 24 months of signal detection. $ACT = n_signals_with_regulatory_action / n_total_detected_signals$. Score 0: $ACT < 10\%$. Score 0.5: ACT 10-30%. Score 1.0: $ACT \geq 30\%$. Actionability reflects both signal quality (true positives have higher ACT) and regulatory follow-through. Two-Source Confirmation Protocol: STEP 1 -- Apply standard SRS disproportionality ($BCPNN\ IC_{025} > 0$ OR $PRR > 2$ AND chi-squared > 4). STEP 2 -- For all SRS signals meeting threshold: query Active Surveillance (Sentinel/DARWIN EU) with equivalent drug-outcome pair as secondary analysis. STEP 3 -- If signal confirmed in Active Surveillance ($OR > 1.3$; $p < 0.05$ in propensity-matched analysis): classify as 'Confirmed Signal'; escalate to regulatory signal committee within 30 days. STEP 4 -- If SRS signal NOT confirmed in Active Surveillance: classify as 'Provisional Signal'; continue enhanced monitoring quarterly; do not escalate unless SRS signal strengthens ($IC_{025} > 1.0$) or clinical case severity increases. STEP 5 -- If NLP-EHR signal detected independently without SRS signal: treat as Layer 4 provisional; query Active Surveillance for confirmation (Steps 2-4 equivalent). EXPECTED OUTCOME: reduction in signal assessment workload of 52.6% (from 47.4% FAERS confirmation rate) with maintenance of sensitivity for true positive signals confirmed in epidemiological analysis.