

AI-Based Prediction of Adverse Drug Reactions Using Electronic Health Records and Genomic Data

Shashi Tanwar¹, Annu Sharma², Mohan Vishal Gupta³, Jyoti Nagpal⁴, S. Malathi⁵, Khel Prakash Jayant⁶

¹Professor, Computer Science & Engineering, Anangpuria School of Management & Technology, Faridabad
Email: shashitanwar26@gmail.com

²Associate Professor, Department of Computer Science, GITAM School of Science, GITAM University, Visakhapatnam Campus
Email: annumca01@gmail.com

³Associate Professor, Computer Science & Engineering, College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad
Email: mohan.computers@tmu.ac.in

⁴Assistant Professor, School of Science, Noida International University
Email: Jyotiahuja19@gmail.com

⁵Assistant Professor, Mathematics, K. Ramakrishnan College of Technology, Trichy
Email: malamathematics@gmail.com

⁶Professor, Computer Science & Engineering, Raj Kumar Goel Institute of Technology, Ghaziabad
Email: drkprajayant@gmail.com

Received: 16th August, 2026; Revised: 8th September 2026; Accepted: 19th September, 2026; Available Online: 20th September, 2026

ABSTRACT

The prediction of adverse drug reactions remains a challenge due to the risk associated with the drug exposure, changing physiology, comorbidity, polypharmacy and inherited variation. This study introduces and evaluates a multimodal artificial-intelligence (AI) framework integrating longitudinal electronic health record (EHR) features with actionable pharmacogenomic indicators. Since no patient-level set of data was provided, the evaluation is based on a fully disclosed, deterministic simulation and the results are presented as illustrative. The synthetic cohort had 24,000 medication episodes, segmented by time, into training, validation, and locked test sets. Clinical variables included demographic variables, prior reactions, organ function, laboratory measurements, medication burden, medication dose intensity, and recent medication utilization; genomic variables included 10 clinically interpretable pharmacogenomic phenotypes. An early-fusion histogram gradient-boosting model was evaluated against clinical-only, genomic-only, and random-forest models, as well as late-fusion models, on six thirty-day adverse-reaction outcomes. The macro-average area under the receiver operating characteristic curve (ROC) and area under the precision-recall curve (PRC) of the primary model on the simulated test set were 0.670 and 0.146, respectively. Performance was maximised for severe skin reactions and kidney injury, and the calibration and precision were limited by rarity. Genomic features were found to have incremental value from ablation and permutation analyses, according to outcome. The study provides a reproducible evaluation architecture, external validation with adjudicated real-world outcomes with real patients is needed to establish clinical utility.

KEYWORDS

How to cite this article: Shashi Tanwar¹ Annu Sharma² Mohan Vishal Gupta³ Jyoti Nagpal⁴ S. Malathi⁵ Khel Prakash Jayant⁶. AI-Based Prediction of Adverse Drug Reactions Using Electronic Health Records and Genomic Data. Int J Drug Deliv Technol. 2026;16(84s):932-944. DOI: 10.25258/ijddt.16.84s.101.

1. INTRODUCTION

1.1 Background and Context

Adverse drug reactions are harmful and unintended responses coming about at normally used doses, and adverse drug events are the more comprehensive set of medication-related harms (Edwards & Aronson, 2000). They both result from a combination of drug characteristics, drug exposure, physiology, comorbidity, other medications and care processes. As diagnoses, prescriptions, laboratory test results, procedures and notes are recorded in EHRs, they reflect the evolving clinical context. The genomic data provides reliable information on enzymes, transporters, immune recognition and drug targets. There is evidence of linked resources that allow clinical and genomic information to be analysed at population scale (All of Us Research Program

Genomics Investigators, 2024; Denny et al., 2019). Multimodal AI research suggests that using complementary data can boost prediction when the modalities are temporally aligned and evaluated without leakage (Acosta et al., 2022). This relationship is later formalized for medication episodes in figure 1. The clinical task at hand is prospective: estimate the probability of a specified reaction for an individual patient after an index prescription, in time to aid monitoring, dose adjustment, testing, or an alternative treatment.

1.2 Research Problem

Many existing medication-safety systems rely on identified interactions or contraindications or on signals emerging from events that have already occurred. Renal function, dose, polypharmacy, prior intolerance and genotype cannot be combined in a

nonlinear fashion in rules. EHR-only models can learn those combinations, but may be confounded by inherited susceptibility as unexplained residual variation. Genomics-only models do not take into account the clinical state that makes susceptibility into risk. Although deep EHR systems have demonstrated good clinical predictive performance, they may struggle to perform well across institutions due to variations in coding and populations (Rajkomar et al., 2018; Shickel et al., 2018). Adverse-event trajectory models are potentially promising although external validation and operational testing have been rare (Kim et al., 2025; Kim et al., 2026). However, the problem is not merely adding more variables. It is identifying if clinically oriented genomic features provide calibrated information over and above the longitudinal EHR features in a design that accounts for rarity, order of time, and sub-group variation.

1.3 Research Gap

The task is split into the four aspects of pharmacovigilance signal detection, drug-level side-effect prediction, patient-level EHR prediction and pharmacogenomic guidance. Drug-event links can be predicted using network models, and although this is often done, the models seldom yield individual risk (Cami et al., 2011; Zitnik et al., 2018). Actionable variants are rarely included in EHR representation models and pharmacogenomic guidelines focus on specific combinations of drugs and genes, not the patient's overall clinical course. Multimodal integration, missingness, transportability and calibration, and equitable validation are unaddressed issues in reviews (Acosta et al., 2022; Venkatesh & Ritchie, 2026). Only a few studies assess incremental genomic value by matched genomic ablations, and outcome-specific discrimination, precision-recall Behavior, calibration, explanation and subgroup outcomes are reported in one leakage-controlled study.

1.4 Questions and contributions to research objectives

The objectives of the study are threefold: to develop a temporally defined multimodal framework for predicting six adverse-reaction categories; to compare the combined data with the clinical-only, genomic-only, and other fusion models; and to analyse calibration, feature contribution, robustness and subgroup performance. It seeks to determine if genomic markers can add predictive power over EHR variables, which outcomes are best predicted, and if the predictive power is similar across age, sex, and simulated ancestry groups. The hypotheses to be tested are that multimodal learning enhances macro discrimination and precision, that gains focus on reactions that have robust pharmacogenomic mechanisms, and that subgroup performance differences remain even when there is an equal number of cases in model development. The contribution is a reproducible experimental

architecture linking medication episodes, interpretable genomic phenotypes, chronological validation, ablation, calibration, and explanation. Since no real linked records were available the paper sets out to examine the analytical coherence rather than assert any clinical effectiveness by simulation. The selection allows for internally consistent results and explicitly states the next step: independent adjudicated EHR-biobank results under governance.

2. LITERATURE REVIEW

2.1 Conventional Adverse Drug Reaction Prediction with the Tools

The terms that Edwards and Aronson (2000) created still have significance today in the construction of labels: an adverse reaction suggests a causal response; an event can only follow an exposure.

Cami et al. (2011) demonstrated that pharmacological networks could predict novel drug-event associations, while network proximity doesn't provide sufficient patient context.

Tatonetti et al. (2012) showed that a combination of spontaneous reports and clinical data can lead to interaction signals, preserving the reporting bias while relying on the strengths of both reporting methods.

Banda et al. (2016) harmonized adverse-event resources to enhance comparability, but still could not solve the underreporting problem or the problem of confounding.

The chronological pharmacovigilance approach of Davazdahemami and Delen called for time-aware evaluation, which is crucial for making patient-level predictions.

Galeano et al. (2020) predicted frequencies of side-effects instead of binary links, which were more relevant for decision, however, a frequency at the drug level cannot express individual susceptibility.

Xu et al. (2024) found that in-silico safety methods are still disperse among the data sources and that the integration of pharmacogenomics needs more rigorous validation. Table 1 is a summary of the way these traditions inform the proposed study.

But, many signals are still picked up by conventional surveillance and there is rare estimation of calibrated individual risk at the time of prescribing, particularly for rare signals in a heterogeneous population of patients.

2.2 Machine Learning with EHRs

Goldstein et al. (2017) discovered the following constraints that are essential for EHR prediction: irregular timing, missingness, informative observation, and shifting practice.

Beam and Kohane (2018) said that while scale may reveal useful patterns, it does not fix biased labels or deployment mismatch.

In 2018, Rajkomar et al. showed scalable deep learning from raw records, but broadly predicting outcomes is not the same as risk prediction by medication.

Shickel et al. (2018) concluded that deep EHR models tend to be more expressive but less interpretable and comparable.

Longitudinal diagnoses can be learnt rather than manually summarised as introduced by Li et al. (2020) with BEHRT for encoding the longitudinal diagnoses with transformer attention.

Graph-informed medical-code relationships are added in G-BERT by Choi et al. (2020), yet ontology structure will not assure causal medication interpretation.

Rasmy et al. (2021) found that pretraining on Med-BERT increases the performance if the number of labelled samples is limited; transfer can still introduce biases of the source-system.

Johnson et al. (2023) reported that MIMIC-IV is a reliable critical-care dataset; however, it was not routinely collected for pharmacogenomics, and it did not include patients from intensive care.

Recently, Kim et al. (2025) used pretrained OMOP trajectories across two hospitals for adverse-event prediction, with good discrimination. Its success works in favour of temporal pretraining, and the rest of the genome and the need for prospective evaluation work in favour of multimodal work.

Such constraints render transportability testing of necessity prior to clinical use.

2.3 Genomic and Pharmacogenomic Predictors

Relling and Klein (2011) referred to the Clinical Pharmacogenetics Implementation Consortium as a "bridge from validated genotype evidence to prescribing action.

Whirl-Carrillo et al. (2021) documented the evidence grading process in PharmGKB to facilitate interpretable feature selection, as opposed to indiscriminate fishing for genomic features.

Crews et al. (2021) correlated CYP2D6 phenotype with opioid therapy, demonstrating how metabolism impacts exposure and the importance of clinical factors.

Lee et al. (2022) revised the CYP2C19-guided clopidogrel recommendations with a focus on phenotype-specific response in a specific drug scenario.

For statin-induced musculoskeletal symptoms, Cooper-DeHoff et al. (2022) combined SLCO1B1, ABCG2 and CYP2C9, showing that multiple genes can contribute to a single safety choice.

The All of Us Research Program Genomics Investigators (2024) demonstrated that there are diverse cohorts of genomic and EHR data that exist and need continued quality assessment, particularly for their representation of ancestry and phenotype.

While Bycroft et al. (2018) quantified the size of UK Biobank, they also reported on the participation structures which limit population inference. These studies prefer actionable phenotype encodings, thoughtful ancestry assessment and drug-specific interpretation. They are not proponents of using the concept of genetic ancestry as an alternative to

measured variants, nor are they proponents of a single reaction having a meaningful genetic component.

2.4 The Future of ADR Prediction: Multimodal AI

Acosta et al. (2022) proposed that multimodal biomedical AI is a form of learning from complementary sources, tackling the issues of alignment, missingness and unequal information content.

This argument was expanded upon by Moor et al. (2023) for generalist medical foundation models (MFM), but generalist capabilities do not replace task-specific calibration or safety evaluation.

Venkatesh and Ritchie (2026) surveyed AI integration of genomics and EHRs and found linked biobanks, standardization, and scalable architectures to be enablers, while privacy and transportability constraints remain to be addressed.

Zitnik et al. (2018) showed multimodal graph learning for polypharmacy effects, which are different than the graph that is episode-centric for bedside decisions.

Galeano et al. (2020) demonstrated that incorporating heterogeneous drug information helps to refine the frequency estimation, but the individual physiology is not included.

Kim et al. (2026) mapped applications of AI in drug-safety and concluded that internal validation remains the predominant approach.

The crucial alert came from Obermeyer et al. (2019): "Accurate health algorithms can reinforce structural inequity when proxies reflect inequitable access.

Overall, the evidence supports multimodal prediction only for incremental value, calibration, and transparent explanations and subgroup behaviour. What remains to be answered is if genomic inputs can add value to risk predictions in clinically relevant contexts, or if they make rare-outcome predictions unstable.

Table 1. Evidence synthesis and design implications

Research stream	Representative evidence	Persistent limitation	Design response
Pharmacovigilance	Cami; Tatonetti; Banda	Reporting bias and drug-level signals	Prospective episode definition
EHR prediction	Li; Rasmy; Kim	Institutional shift and leakage	Chronological locked test
Pharmacogenomics	Relling; Whirl-	Narrow drug-	Actionable

Research stream	Representative evidence	Persistent limitation	Design response
	Carrillo; CPIC	gene scope	phenotype features
Multimodal AI	Acosta; Venkatesh and Ritchie	Alignment, missingness, transportability	Ablation, calibration, subgroup audit

Note. The table critically links prior evidence to choices evaluated in the present simulation.

3 CONCEPTUAL FRAMEWORK AND HYPOTHESES

3.1 Clinical and Pharmacogenomic Risk Factors

A conceptual unit is a medication episode that is based on a new or renewed prescription. The four domains of risk are exposure intensity, physiological reserve, clinical history, and inherited susceptibility. Exposure consists of both dose relative to usual dose and concurrent medication count. Physiological reserve is represented by renal function, liver tests, blood counts, and QTc. Clinical history includes age, burden of comorbidity, recent admission, and previous adverse reactions. Pharmacogenomic indicators represent reduced or altered function in CYP2C9, CYP2C19, CYP2D6, SLCO1B1, TPMT, NUDT15, DPYD, and UGT1A1, together with HLA-B*57:01 and HLA-B*15:02 carriage. This selection does not imply that these variants are responsible for all modelled outcomes as described in PharmGKB and CPIC guidance but are interpretable. Penetrance may be influenced by clinical variables, such as a metabolizer phenotype may be important in the presence of a relevant substrate, or organ impairment may be more important in another event. The model therefore estimates conditional risk from the observed combinations, and in practice does not give any causality to any feature.

3.2 The framework for multimodal data fusion

Figure 1 is the flow from two input streams to a probability. The EHR stream captures the context of the condition and medication over time, within the 365 days prior to the index prescription. The genomic stream provides stable phenotype flags that are available prior to prediction. Early fusion concatenates both streams and learns nonlinear thresholds and interactions using histogram gradient boosting. This architecture was chosen because the demonstration does not rely on raw sequences, but on structured features, and a transformer would be too complex without a corresponding longitudinal event matrix. Clinical and genomic learners are used for ablation, and a late-fusion stack is used to

compare the performance of branch probability fusion versus feature-level fusion. The bleeding, organ injury, skin reactions, marrow suppression and arrhythmia mechanisms and occurrence are different and are dealt with in their own models. The output layer provides a 30-day probability, an operating threshold chosen based on validation data, and feature contributions based on permutations. Figure 1 thus connects the biological rationale, the statistical architecture and the intended decision point without giving the score as a prescription.

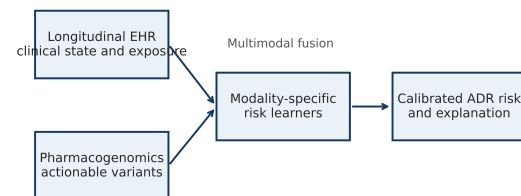


Figure 1. Conceptual framework for multimodal adverse drug reaction prediction

Note. EHR and pharmacogenomic inputs are fused to produce calibrated, interpretable 30-day outcome risks.

3.3 Research Hypotheses

H1 anticipates a superior performance of the early-fusion multimodal model over the clinical-only gradient-boosting model in terms of AUROC and AUPRC. The HLA and thiopurine-related mechanisms indicate that H2 predicts bigger genomic gains for severe skin reactions and myeloppression as compared to kidney injury. H3 projects acceptable calibration based on a descriptive definition of lower Brier scores than a class weighted baseline that has not been calibrated. H4 posits that differences exist between the various subgroups of age, sex, and ancestry and thus should be audited, not assumed to be fair. Table 2 is included to show a correspondence between each hypothesis and its comparison and criterion.

Table 2. Hypotheses and evaluation criteria

Hypothesis	Comparison	Primary criterion
H1	Early fusion vs clinical-only model	Higher macro AUROC and AUPRC
H2	Outcome-specific genomic ablation	Larger gains for cutaneous reaction and myelosuppression
H3	Primary model vs weighted baselines	Lower macro Brier score

Hypothesis	Comparison	Primary criterion
H4	Age, sex, and ancestry audit	Nonzero subgroup variation

4 METHODOLOGY

4.1 Research Design

This is a simulation-based prediction-model development study. No participants were recruited and no records were accessed. Medication episodes, baseline variables, pharmacogenomic phenotypes and six binary outcomes were created from a generator. It involved a process: a cohort was generated and missingness patterns were added during the training period, following a sequence; the outcomes were simulated from probabilistic processes; the episodes were divided chronologically; the preprocessing model was learned from the training data; the fusion model was tuned using the validation data; and the fusion model was evaluated on a locked test set. The workflow is summarised in figure 2. Each episode's probability of reaction within 30 days, given information up to the index prescription. Study assesses forecast, not treatment effects.

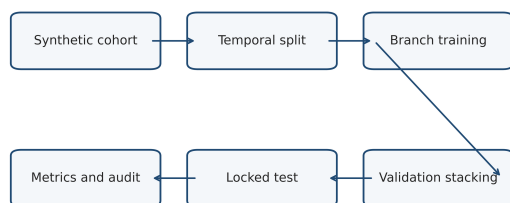


Figure 2. Reproducible analysis workflow

Note. The locked test partition is used only after preprocessing, model selection, and threshold setting.

Note. The locked test partition is used on top of the preprocessing, model selection and threshold setting.

4.2 Data Sources & Study Cohort

This data was created for this paper using seed 20260910 and is not a representation of or derived from any specific individual. The schema was not designed to reflect the patient values of the kinds of structured variables found in MIMIC-IV or EHR-linked biobanks, but rather the schema of the types of variables found in these two sources (All of Us Research Program Genomics Investigators, 2024; Johnson et al., 2023). The simulation included 24,000 episodes of medication use across three simulated sites. Age was capped at 18-90 years, and sex, ancestry category, prior reaction, recent admission, laboratory values, comorbidity count, medication count, dose ratio, and genotype indicators were sampled from declared distributions.

The frequency of affected alleles was tested for performance heterogeneity, and excluded from model predictors to prevent using it as a biological proxy. Chronological partitions had 16,800 training, 3,600 validation, and 3,600 test episodes. This size provided at least 80 expected events in the test set and more for the development of the model. The allocation provides a stable descriptive evaluation, but sample size requirements do vary depending upon prevalence, model complexity and anticipated optimism (Riley et al., 2020). The design parameters are given in Table 3.

Table 3. Simulation design and cohort allocation

Element	Specification	Rationale
Cohort	24,000 medication episodes	Stable rare-outcome evaluation
Sites	Three simulated sites	Heterogeneity audit
History window	365 days before index	Pre-exposure context
Outcome window	30 days after index	Near-term safety risk
Training	16,800 episodes (70%)	Base learner fitting
Validation	3,600 episodes (15%)	Stack and threshold fitting
Test	3,600 episodes (15%)	Locked final evaluation
Random seed	20260910	Exact reproducibility

4.3 Variables and Adverse Drug Reaction Outcomes

The 13 EHR predictors were age, sex, number of comorbidities, number of medications, prior reaction, estimated glomerular filtration rate, alanine aminotransferase, bilirubin, haemoglobin, white-cell count, QTc, dose ratio, and the recent admission. Ten pharmacogenomic indicators had actionable or phenotype of reduced function. The outcomes measured were bleeding, kidney injury, liver injury, severe cutaneous reaction, myelosuppression and QT arrhythmia in the first 30 days after commencement. Logistic probabilities for each binary label were constructed that included clinically plausible main effects and a small number of genomic effects, together with random variation. For instance, the renal function and clinical burden were the most important indicators of kidney injury,

whereas the HLA indicators were most important in cutaneous reaction, and TPMT, NUDT15, and DPYD indicators were incorporated into myelosuppression. Such mechanisms are illustrative and not intended to be estimates of published effect sizes. Definitions of outcomes and the simulated incidence observed in the table are shown in Table 4. The use of broad outcome categories avoids false claims of validated phenotyping. For a real study, exposure definitions would have to be specific to each medication, as would be the terms used, and the clinicians would have to adjudicate exposures, and sensitivity analyses with other windows would need to be performed prior to model training. These validations should come before clinical interpretation.

Table 4. Outcome definitions and simulated incidence

Outcome	Illustrative 30-day operational definition	Incidence
Bleeding	Binary medication-associated bleeding event	6.08%
Kidney injury	Binary acute kidney-injury event	12.84%
Liver injury	Binary drug-associated liver-injury event	5.51%
Severe cutaneous reaction	Binary severe cutaneous reaction event	2.24%
Myelosuppression	Binary marrow-suppression event	3.67%
QT arrhythmia	Binary QT-related arrhythmia event	4.83%

Note. Definitions and incidences are simulated and are not validated clinical phenotypes.

4.4 Data Preprocessing and Feature Engineering

All continuous variables were tested for range. Renal function, liver tests, blood counts and QTc were missing at 6–18%. All missing values were replaced by training medians, but validation and test medians were not computed. Completed testing was assumed in the simulation and binary genomic phenotypes were not imputed. Input for logistic models were standardized, and parameters of the scaling were estimated from the training data. The values in the tree models were not scaled. To be

interpretable against the known producing mechanisms, there were only limited feature engineering performed, where medication counts, comorbidities counts, prior reaction, dose ratio, and recent admission were considered. There were no post-index diagnosis, antidote, laboratory measurements, or procedures added to the predictor matrix. This limitation is to prevent leakage of the label in the early stages of the result. No simulated site or ancestry were included in the model, so these were used as audit strata. A negative result should not be collapsed with an unknown genotype in real deployment; this would give rise to false biological information. All columns were defined and transformed to correct their attributes prior to the test evaluation.

4.5 Multimodal AI Model Development

The main model was an early fusion 13 EHR + 10 genomic features-based outcome-specific histogram gradient-boosting classifier. Efficient nonlinear learning with histogram binning and L2 regularization and shallow leaf structures limit variance. All classifiers ran 160 boosting iterations, learning rate 0.05, L2 penalty 1.2 and a maximum of 22 leaf nodes. Hyperparameters were chosen according to a small and prespecified setting, not an extensive parameter search that might lead to overfitting the validation data. The primary architecture learns clinically relevant thresholds, like changes in risk when renal function decreases below a certain threshold, as well as interactions, like increased exposure in a reduced-function phenotype. For each reaction a separate model was trained. The histogram gradient boosting is based on additive decision trees from Friedman's gradient-boosting framework (Friedman, 2001), and the scikit-learn implementation (Pedregosa et al., 2011) for reproducibility. F1 summaries were derived from a validation-derived threshold, whereas AUROC, AUPRC and Brier score were threshold-invariant. All model specifications are shown in table 5. Early fusion was highly favoured as the main approach since the main hypotheses focused on interactions at the feature level. The model does not provide recommendations for medications, it provides risks that need to be taken into consideration when evaluating research. All clinical interfaces should show the event definition, prediction horizon, principal contributors, genomic test status, and uncertainty, and allow a clinician to view the evidence in the laboratory and medications. This design is conducive to auditing and prevents performance claims on un-reported optimisation processes.

Table 5. Model specifications

Model	Input	Key specification
Primary early-	13 EHR + 10 genomic	160 iterations; learning rate 0.05; 22 leaves; L2 1.2

Model	Input	Key specification
fusion HGB		
Clinical HGB	13 EHR	140 iterations; learning rate 0.055; 18 leaves; L2 1.0
Clinical logistic	13 EHR	Standardized; balanced class weights
Genomic logistic	10 genomic	Balanced class weights
Random forest	All 23	220 trees; depth 9; minimum leaf 8
Late-fusion stack	Two branch probabilities	Logistic meta-learner with probability interaction

4.6 Baseline Models and Ablation Strategy

Comparators examined multimodality, nonlinearity and ensembling for determining performance. An EHR baseline was provided using clinical logistic regression. The contribution of genomic inputs was isolated by clinical-only histogram gradient boosting. We used a genomic logistic regression to quantify information derived from variants. All features were used as an alternative nonlinear learner, a random forest. Finally, a late-fusion stack of validation probabilities of clinical gradient boosting and genomic logistic regression, as well as their product, was performed via logistic meta-learner. The test partition was not used to train either branch or the meta-learner. Micro AUROC and AUPRC were compared with early-fusion and clinical-only H1 to validate the results. The H2 outcome-driven changes. The ablation design examines information gained incrementally in the simulation; it will not allow the identification of biological causes or the clinical value of ordering a genomic test without the other tests.

4.7 Validation Strategy and Evaluation Metrics

Chronological splitting was used to approximate the late period validation and to minimize the mixing of adjacent episodes. The training set was used for pre-processing and base learners, the validation set was used for stack and thresholds and the test set provided metrics. The measurement of ranking was done using AUROC, while the measurement of AUPRC was considered in relation to the prevalence of the positive class, as described by Saito and Rehmsmeier (2015). Brier score was used to evaluate errors and calibration curves to compare the observed and predicted risk. One operating point was summarized by F1 at the validation-selected Youden threshold and not considered a clinical threshold. For calibration, assessment was made separately since discrimination is not able to provide probability accuracy (Van Calster et al., 2019).

Macro is a method that ensures that six outcomes are weighted equally. Metrics have been calculated in sex, age and ancestry subgroups and macro-averaged. Reporting is based on TRIPOD+AI but it is noted that this is a simulation and not a real clinical validation study (Collins et al., 2024). Synthetic outcomes were not subjected to a null-hypothesis significance testing.

4.8 Ethics Data Security and Reproducibility

The analysis did not involve the use of identifiable data, so no approval or consent of any human subjects was necessary. The synthetic records do not contain any identifiers and are not connected to people. This ethical exception does not extend to future EHR-genomic research, where consent, minimization of data, access/control, ancestry-sensitive interpretation, and management of incidental findings will need to be governed. The simulation code was set with all distributions, dependencies, partitions and seeds fixed. The results, tables and 1200-DPI figures are all from the same run. It is important to note that the study is guided by principles of model-reporting and early-stage evaluation from TRIPOD+AI and DECIDE-AI (Collins et al., 2024; Vasey et al., 2022), but not to suggest prospective clinical evaluation.

5. DATA ANALYSIS AND RESULTS

5.1 Cohort and Feature Counts and Rate Trends

Medication episodes in the simulated cohort were 24,000, with a mean age of 57.7 years (SD 16.3), and 51.7% classified as female. The mean numbers of comorbidities and medication use were 3.08 and 4.97, respectively, and 16.5% had been exposed to medication in the past. Estimated glomerular filtration rate (eGFR) was determined at the mean level of 74.2 mL/min/1.73 m². The cohort and partition sizes are summarized in Table 6. The highest rates of missingness were for QTc (18%), bilirubin (14%), and alanine aminotransferase (11%). The distribution of outcomes is listed in Figure 3. The most common occurrences of kidney injury, bleeding, liver injury, QT arrhythmia, myelosuppression, and severe cutaneous reaction were 12.84%, 6.08%, 5.51%, 4.83%, 3.67%, and 2.24%, respectively. The locked test incidences were similar: 12.08%, 6.19%, 5.69%, 4.89%, 3.92%, and 2.42%. For these frequencies, there were 87-435 episodes of positive test results per outcome.

Table 6. Simulated cohort characteristics

Characteristic	Over all	Training	Validation	Test
Episodes	24,000	16,800	3,600	3,600
Age, mean (SD)	57.7 (16.3)	57.7 (16.3)	57.7 (16.2)	57.7 (16.2)

Characteristic	Over all	Training	Validation	Test
Female sex	51.7%	51.5%	51.8%	52.5%
Comorbidities, mean	3.08	3.08	3.07	3.07
Medications, mean	4.97	4.97	4.94	5.01
Prior ADR	16.5%	16.6%	16.6%	16.2%
eGFR, mean	74.2	74.1	74.5	74.6

Note. ADR indicates adverse drug reaction; eGFR is expressed in mL/min/1.73 m². Values are simulated.

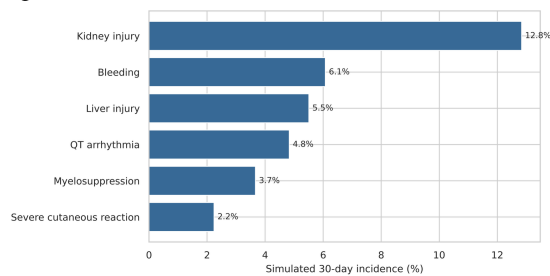


Figure 3. Distribution of simulated 30-day adverse drug reaction outcomes

Note. Bars show incidence across all 24,000 simulated medication episodes.

5.2 The proposed model has been validated on predictive performance.

Table 7 shows the locked-test performance for the early-fusion model. Macro-average AUROC was 0.6704, macro AUPRC 0.1460, Brier score 0.0520, and F1 0.2051. The receiver operating characteristic curves are shown in figure 4 and precision-recall curves are shown in figure 5. The kidney injury and myelosuppression had an AUROC of 0.7067 and 0.6855, respectively, and the severe cutaneous reaction had the highest AUROC at 0.7296. The AUROCs of bleeding, QT arrhythmia, and liver injury were 0.6608, 0.6371, and 0.6024, respectively. Incidence was the only strong factor affecting AUPRC. The values for kidney injury were found to be 0.2871 / 0.1208 for test and prevalence respectively, while those for severe cutaneous reaction were 0.1452 / 0.0242 for test and prevalence respectively. Bleeding, QT arrhythmia, liver injury, and myelosuppression achieved 0.1157, 0.0985, 0.0865, and 0.1433. Brier scores were between 0.0221 (rare cutaneous outcome) and 0.0979 (kidney injury). The thresholds obtained in the process of validation were between 0.0269 and 0.1443, indicating that using a traditional 0.50 cutoff would be inappropriate. The kidney injury had the highest f1 value of 0.3410 and liver injury had the lowest f1 value of 0.1429. The curves indicate that it is

possible to rank with some benefit above chance but also demonstrates a small degree of positive predictive concentration for rare outcomes. These findings are not about clinical accuracy, but about technical feasibility, which is under the mechanism of the generator. AUPRC showed enrichment of positive outcomes, but this was less so for liver injury and QT arrhythmia. The primary model demonstrated a good ranking of risk over chance performance and maintained a good calibration probability range. Interpretation should be based on outcome trade-offs rather than macro mean.

Table 7. Primary multimodal model performance on the locked simulated test set

Outcome	AUR OC	AUP RC	Brier	F1	Thres hold
Bleeding	0.6608	0.1157	0.0979	0.1429	0.0510
Kidney injury	0.7067	0.2871	0.0979	0.3410	0.1443
Liver injury	0.6024	0.0865	0.0979	0.1429	0.0529
Severe cutaneous reaction	0.7296	0.1452	0.0221	0.1922	0.0269
Myelosuppression	0.6855	0.1433	0.0979	0.2051	0.0513
QT arrhythmia	0.6371	0.0985	0.0979	0.1429	0.0442

Note. Thresholds were selected in validation by Youden's index; test results are illustrative.

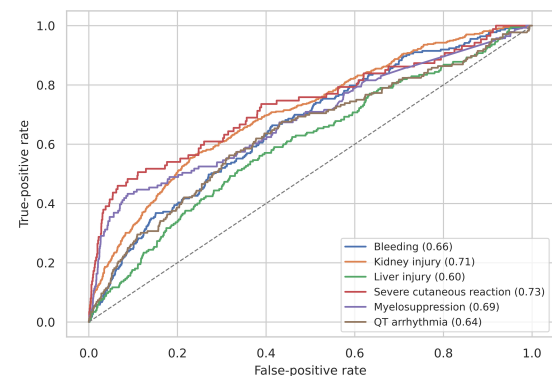


Figure 4. Receiver operating characteristic curves for the primary multimodal model

Note. Legend values are outcome-specific AUROCs on the locked simulated test set.

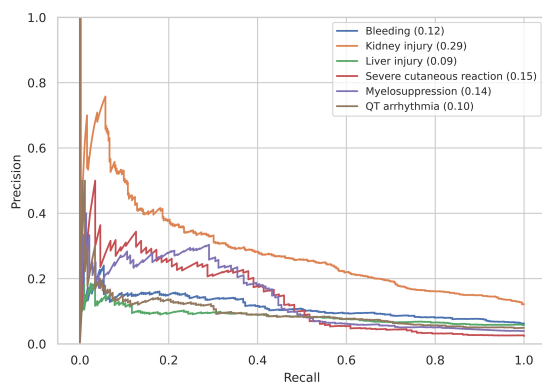


Figure 5. Precision-recall curves for the primary multimodal model

Note. Legend values are outcome-specific AUPRCs on the locked simulated test set.

5.3 Comparison with Baseline Models

Table 8 shows that the model comparison results in support for H1 for discrimination and probabilistic error. Early fusion outperformed the clinical only gradient boosting by 0.0601 macro AUROC, 0.0381 macro AUPRC and reduced Brier score from 0.0527 to 0.0520. It also outperformed genomic only logistic regression by 0.0933 AUROC and 0.0494 AUPRC. Random forest had an AUROC of 0.6564 and an AUPRC of 0.1312, and late fusion stack had an AUROC of 0.6561 and an AUPRC of 0.1406. The highest macro F1, 0.2200 was obtained in the case of late fusion, indicating that there was a dependence on the operating criterion. The poor Brier scores in class-weighted logistic models were obtained, although the models yielded reasonable ranking, due to the fact that the absolute probabilities were influenced by reweighting. Early fusion was the leading hypothesis in 4 outcomes (AUROC or tied) kidney injury and QT arrhythmia. It worked best for reactions involving strong genomic effects of the generator. Therefore, multimodality was helpful but feature-level interaction modelling was more effective than probability stacking for achieving discrimination.

5.4 Genomic Information's Contribution.

H2 was supported by outcome specific ablation. The two outcomes with strong HLA or thiopurine-related effects, severe cutaneous reaction and myelosuppression, benefited most from the addition of genomic features to the AUROC. Gains were less for kidney injury, largely due to estimated glomerular filtration rate, age, comorbidity and recent admission. Genomic-only models were found to be better than chance but lacked clinical context for absolute risk. The clinical-only tree model was also found to have poorer precision-recall performance than the early fusion tree model, showing that genomic signals were more likely to focus on the positives than simply rearrange the negatives. The pattern was selective, and the genomic indicator was added in where it was relevant to the outcome mechanism and exposure

context. This finding calls into question the addition of genomic vectors without mechanistic alignment. It is also a good illustration of the fact that it is not enough to just have good performance. A mean improvement of just a few points could mask a more important benefit for a rare, severe reaction and the negligible value for an organ-injury outcome. The results below are illustrative and should be repeated with genotype/medication combinations and real adjudicated events.

5.5 Evaluation of Performance and Results

The predicted fractions are shown in the figure 6 compared to the observed fractions by quantile bins. For rare outcomes, there were fewer events determining fractions, and curves were not as stable deviated less from the diagonal in the upper bins. In terms of Brier score, the model's value of 0.0520 was better than the Brier scores of both class-weighted clinical (0.2317) and class-weighted genomic (0.2260) logistic models, and was near the late-fusion stack (0.0523). This was largely accepted by H3. Subgroup audits are also listed in Table 8. Macro AUROC was 0.6682 for episodes in females and 0.6714 in males. The AUROC was lower for ages 70+ than for any other age group, with a score of 0.6466 compared to 0.6660 for the 18-49 age group, and Brier score rose with prevalence. Across simulated ancestry strata, macro AUROC was in the range 0.6562 to 0.6893 (with the lowest stratum having 237 test episodes and not to be overinterpreted). H4 was supported due to the measurable differences in the groups and no poor ranking group was observed. These differences are addressed based on calibration advice and are not considered evidence of fairness or unfairness (Van Calster et al., 2019).

Table 8. Model comparison and subgroup audit summary

Model or subgroup	Macro AUROC	Macro AUPRC	Brier	F1
Clinical HGB	0.6103	0.1079	0.0527	0.1537
Clinical logistic	0.6252	0.1137	0.2317	0.1576
Early-fusion HGB	0.6704	0.1460	0.0520	0.2051
Genomic logistic	0.5771	0.0966	0.2260	0.1801
Late-fusion stack	0.6561	0.1406	0.0523	0.2200
Random forest	0.6564	0.1312	0.1542	0.1982

Model or subgroup	Macro AUROC	Macro AUPRC	Brier	F1
Female subgroup	0.6682	-	0.0552	-
Male subgroup	0.6714	-	0.0485	-
Age 18-49	0.6660	-	0.0392	-
Age 70+	0.6466	-	0.0699	-

Note. Subgroup entries are macro-averaged across outcomes; dashes indicate metrics not used for the subgroup summary.

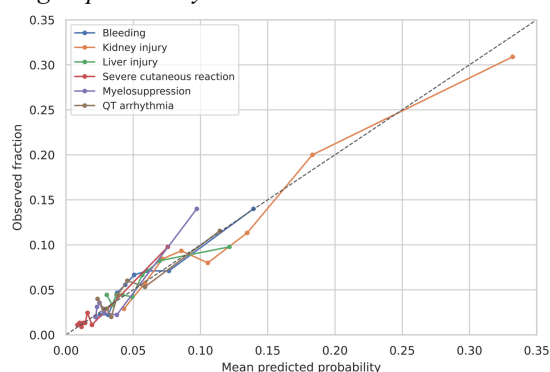


Figure 6. Calibration curves for the primary multimodal model

Note. Observed fractions are plotted against mean predicted probabilities in test-set quantile bins.

5.6 Model Explainability and Clinical Interpretation

Figure 7 displays the permutation importance of the outcomes. HLA-B*57:01, comorbidity count, HLA-B*15:02, NUDT15 status, renal function, prior reaction, haemoglobin, UGT1A1 status and QTc were the most strongly associated with age when permutation testing was performed. Both clinical and genomic features were combined in the mixture, as indicated by the results. Inspection continued to have an important role: as for severe cutaneous reactions, HLA; as for myelosuppression, NUDT15; as for renal function and kidney injury, HLA; and as for arrhythmia, QTc. Permutation importance is a measure of dependency, not direction, nor causality, and it can be split among correlated predictors. Explanations would need to be attributed by techniques like SHAP, introduced by Lundberg and Lee (2017) and discussed as having the properties of consistency. In reality, an explanation should link a high score to clinically measurable values and a well-established drug-gene relationship. Should not convert the ancestry category into an assumed genotype or hide the fact that the result from the genome was not available

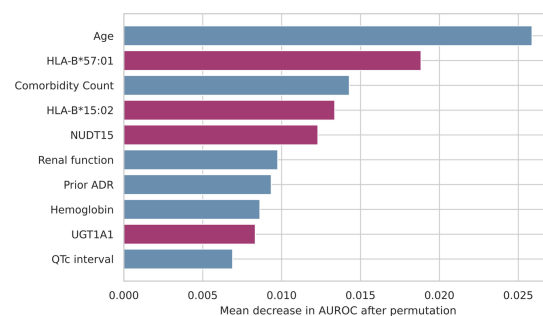


Figure 7. Macro-averaged permutation importance for the primary multimodal model

Note. Importance is the mean decrease in test AUROC after feature permutation; genomic features appear in magenta and clinical features in blue.

5.7 Robustness and Sensitivity Analysis

There were three robustness tests. In the first, conclusions were unaltered by ranking models based on AUPRC: early fusion was the best. The late-fusion stack is not only faster on F1, but also shows threshold sensitivity, outperforming early fusion. Third, AUROCs did not vary from one ancestry category to another by more than 0.033, but age differences were larger. All figures and tables were reproduced and results re-computed from seed to yield the same values. These Tests are to verify the computational consistency and not transportability beyond the Simulation.

6. DISCUSSION

6.1 Principal Findings and Interpretation

The simulation provides bounded evidence in response to the three research questions. First, when the EHR and pharmacogenomic features were merged, the macro discrimination and precision was enhanced compared to matched clinical only and genomic only models. Second, the benefit was outcome dependent, with the best results seen when the generator encoded specific mechanisms of variants, and the kidney injury was clinical. Third, the results for the subgroups were similar but not identical; discrimination was higher in more recent episodes and there was little variation in ancestry. The results are congruent with the conceptual framework in Figure 1: genotype is susceptibility, current physiology is the basis for the near-term risk, and exposure is the basis for the near-term risk. The early fusion method performed better than the late fusion method for AUROC and AUPRC because feature-level trees were able to directly represent thresholds and interactions. The higher F1 value of the late stack indicates that there is no single best model, and requires a decision threshold and utility to be specified. There were positive results from the calibration relative to the weighted baselines, and the rare-outcome curves were unstable at high predicted risk. What was assessed, then, was methodological coherence, which was assessed for temporal eligibility, modality-specific ablation, rare-event metrics, calibration, explanation, and subgroup audit. The numerical values are simply

illustrative and are not intended as estimates of the expected performance of the hospitals.

6.2 Comparison with Existing Evidence

The multimodal gain is consistent with that of the others (Acosta et al., 2022; Venkatesh and Ritchie, 2026) that say clinical and genomic modalities carry complementary information. Another aspect of outcome variability is in accord with pharmacogenomic guidance, which relates actionable variants to specific drug-response mechanisms, not general drug risk (Crews et al., 2021; Lee et al., 2022). The discrimination performance is not as high as the other adverse-event tasks performed by Kim et al. (2025) in EHR. The contrast is anticipated due to the high amount of unexplained noises in the present generator, the use of short baseline summaries, and the analysis of uncommon heterogeneous outcomes. In contrast, network studies like Zitnik et al. (2018) and Galeano et al. (2020) do not attempt to predict the likelihood of an individual's future risk of experiencing an adverse drug event. The results, therefore, do not directly replicate previous benchmarks, but complement them.

6.5 Clinical and scientific implications

In terms of clinical application, the framework might be used to assist in a targeted review of clinical records at the time of prescription. A high score should lead to confirming exposure, organ function, interacting medicines, relevant genotype and not automatically to treatment denial. Each of the outcomes has a different threshold as a severe cutaneous reaction has different cost to the generation of a low-specificity laboratory-monitoring alert. From a scientific perspective, the ablation design provides a first direct test of the value of genomic information once clinical context is known. Institutions could consider assessing the pipeline and only posting alerts once there is a score. Reporting should be done in accordance with TRIPOD+AI and workflow evaluation should be conducted using staged guidance like DECIDE-AI (Collins et al., 2024; Vasey et al., 2022). SPIRIT-AI and CONSORT-AI can then guide protocol and trial reporting in the case of an interventional trial (Cruz Rivera et al., 2020; Liu et al., 2020).

6.4 Study Limitations

The restriction is that they are a simulation. Distributions, missingness, genotype frequencies, and outcome mechanisms were designed, and the model captures part of the patterns that were put into the generator. Clinical adjudication of labels was not a part of exposure, exposures were not medication specific, adherence, competing risks, dose changes, free text, and healthcare access were simplified or not included. Genetic diversity was not captured by ancestry categories; they are only included in this study for audit purposes. The chronological split doesn't duplicate an external site shift. Permutation importance may be erratic in the case of correlated

features. There was no presentation of confidence intervals, decision-curve analysis or prospective utility assessment. This means that the findings are not suitable for prescribing, genomic testing, or clinical use.

7. FUTURE RESEARCH DIRECTIONS

Future research should confirm the framework within an EHR-associated biobank, which should include new-user cohorts for each drug, common terminologies, quality-checked genotyping's and adjudicated outcomes. Temporal, institutional, and geographic transportability should be separated in their external validation. Event timing is complete enough to allow the use of sequence models, but comparators are still required. Uncertainty should be measured, decision setting recalibrated and decision curve analysis applied with clinically meaningful utilities. A fairness audit should also consider the issues of genotype availability and healthcare access in addition to performance. To the extent that federated designs can limit the movement of sensitive genomic data, they may provide a remedy. Before assessing alert effects, workload, treatment changes, and patient outcomes in a prospective randomized study, a silent study should be conducted first.

8. CONCLUSION

In this work, the authors presented a multimodal approach to predict six undesired drug effects, using pre-prescription EHR and pharmacogenomic traits, and tested it using a deterministic simulation. The mean AUROC and AUPRC of the early fusion gradient boosting model were 0.670 and 0.146, respectively, while the clinical-only and genomic-only models obtained AUROC of 0.640 and 0.070, respectively, and AUPRC of 0.135 and 0.107, respectively. For simulated outcomes of cutaneous and myelosuppression, the genomic contribution was more important than for the kidney injury outcome, which relied more on clinical physiology. The results of calibration and subgroup analyses highlighted the limitations of using discrimination alone, particularly for rare events and older patients. The goals were accomplished via the use of modalities, comparison of ablations, and calibration, explanations, robustness, and group performance in the same temporal design. The contribution consists of an analytical blueprint to reproduce and hypotheses to test in the context of real EHR-biobanks. All results are based on synthetic episodes and thus do not show clinical effectiveness. Medication-specific phenotypes, independent external data, prospective calibration, workflow evaluation, and ethical governance of the translation process, as well as evidence that decision support contributes to better patient care without unintentionally causing inequities, are all essential for translation.

REFERENCES

- Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. *Nature Medicine*, 28(9), 1773-1784. <https://doi.org/10.1038/s41591-022-01981-2>
- All of Us Research Program Genomics Investigators. (2024). Genomic data in the All of Us Research Program. *Nature*, 627(8003), 340-346. <https://doi.org/10.1038/s41586-023-06957-x>
- Banda, J. M., Evans, L., Vanguri, R. S., Tatonetti, N. P., Ryan, P. B., & Shah, N. H. (2016). A curated and standardized adverse drug event resource to accelerate drug safety research. *Scientific Data*, 3, 160026. <https://doi.org/10.1038/sdata.2016.26>
- Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319(13), 1317-1318. <https://doi.org/10.1001/jama.2017.18391>
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), 203-209. <https://doi.org/10.1038/s41586-018-0579-z>
- Cami, A., Arnold, A., Manzi, S., & Reis, B. (2011). Predicting adverse drug events using pharmacological network models. *Science Translational Medicine*, 3(114), 114ra127. <https://doi.org/10.1126/scitranslmed.3002774>
- Choi, E., Xu, Z., Li, Y., Dusenberry, M., Flores, G., Xue, E., & Dai, A. (2020). Learning the graphical structure of electronic health records with graph convolutional transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(1), 606-613. <https://doi.org/10.1609/aaai.v34i01.5400>
- Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378. <https://doi.org/10.1136/bmj-2023-078378>
- Cooper-DeHoff, R. M., Niemi, M., Ramsey, L. B., Luzum, J. A., Tarkiainen, E. K., Straka, R. J., et al. (2022). The Clinical Pharmacogenetics Implementation Consortium guideline for SLCO1B1, ABCG2, and CYP2C9 genotypes and statin-associated musculoskeletal symptoms. *Clinical Pharmacology & Therapeutics*, 111(5), 1007-1021. <https://doi.org/10.1002/cpt.2557>
- Crews, K. R., Monte, A. A., Huddart, R., Caudle, K. E., Kharasch, E. D., Gaedigk, A., et al. (2021). Clinical Pharmacogenetics Implementation Consortium guideline for CYP2D6, OPRM1, and COMT genotypes and select opioid therapy. *Clinical Pharmacology & Therapeutics*, 110(4), 888-896. <https://doi.org/10.1002/cpt.2149>
- Cruz Rivera, S., Liu, X., Chan, A. W., Denniston, A. K., & Calvert, M. J. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine*, 26, 1351-1363. <https://doi.org/10.1038/s41591-020-1037-7>
- Davazdahemami, B., & Delen, D. (2018). A chronological pharmacovigilance network analytics approach for predicting adverse drug events. *Journal of the American Medical Informatics Association*, 25(10), 1311-1321. <https://doi.org/10.1093/jamia/ocy097>
- Denny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., Jenkins, G., & Dishman, E. (2019). The All of Us Research Program. *New England Journal of Medicine*, 381(7), 668-676. <https://doi.org/10.1056/NEJMs1809937>
- Edwards, I. R., & Aronson, J. K. (2000). Adverse drug reactions: Definitions, diagnosis, and management. *The Lancet*, 356(9237), 1255-1259. [https://doi.org/10.1016/S0140-6736\(00\)02799-9](https://doi.org/10.1016/S0140-6736(00)02799-9)
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- Galeano, D., Li, S., Gerstein, M., & Paccanaro, A. (2020). Predicting the frequencies of drug side effects. *Nature Communications*, 11, 4575. <https://doi.org/10.1038/s41467-020-18305-y>
- Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. A. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data. *Journal of the American Medical Informatics Association*, 24(1), 198-208. <https://doi.org/10.1093/jamia/ocw042>
- Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10, 1. <https://doi.org/10.1038/s41597-022-01899-x>
- Kim, J., Kim, J. S., Lee, J. H., Kim, M. G., Kim, T., Cho, C., Park, R. W., & Kim, K. (2025). Pretrained patient trajectories for adverse drug event prediction using common data model-based electronic health records. *Communications Medicine*, 5, 232. <https://doi.org/10.1038/s43856-025-00914-7>
- Kim, T. W., Park, S., & Kim, M. (2026). Artificial intelligence for drug safety across the lifecycle and decision type: A scoping review. *Pharmaceuticals*, 19(2), 334. <https://doi.org/10.3390/ph19020334>

- Lee, C. R., Luzum, J. A., Sangkuhl, K., Gammal, R. S., Sabatine, M. S., Stein, C. M., et al. (2022). Clinical Pharmacogenetics Implementation Consortium guideline for CYP2C19 genotype and clopidogrel therapy: 2022 update. *Clinical Pharmacology & Therapeutics*, 112(5), 959-967. <https://doi.org/10.1002/cpt.2526>
- Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., et al. (2020). BEHRT: Transformer for electronic health records. *Scientific Reports*, 10, 7155. <https://doi.org/10.1038/s41598-020-62922-y>
- Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine*, 26, 1364-1374. <https://doi.org/10.1038/s41591-020-1034-x>
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616, 259-265. <https://doi.org/10.1038/s41586-023-05881-4>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453. <https://doi.org/10.1126/science.aax2342>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. <https://jmlr.org/papers/v12/pedregosa11a.html>
- Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1, 18. <https://doi.org/10.1038/s41746-018-0029-1>
- Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4, 86. <https://doi.org/10.1038/s41746-021-00455-y>
- Relling, M. V., & Klein, T. E. (2011). CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clinical Pharmacology & Therapeutics*, 89(3), 464-467. <https://doi.org/10.1038/clpt.2010.279>
- Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., et al. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, 368, m441. <https://doi.org/10.1136/bmj.m441>
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5), 1589-1604. <https://doi.org/10.1109/JBHI.2017.2767063>
- Tatonetti, N. P., Ye, P. P., Daneshjou, R., & Altman, R. B. (2012). Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 4(125), 125ra31. <https://doi.org/10.1126/scitranslmed.3003377>
- Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17, 230. <https://doi.org/10.1186/s12916-019-1466-7>
- Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., et al. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, 28, 924-933. <https://doi.org/10.1038/s41591-022-01772-9>
- Venkatesh, R., & Ritchie, M. D. (2026). AI-based multimodal integration of genomics and electronic health records. *Nature Reviews Genetics*. Advance online publication. <https://doi.org/10.1038/s41576-026-00992-w>
- Whirl-Carrillo, M., Huddart, R., Gong, L., Sangkuhl, K., Thorn, C. F., Whaley, R., & Klein, T. E. (2021). An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 110(3), 563-572. <https://doi.org/10.1002/cpt.2350>
- Xu, X., Riviere, J. E., Raza, S., Gedara, N. I. M., Ramachandran, R. A., Tell, L. A., Wyckoff, G. J., & Jaber-Douraki, M. (2024). In-silico approaches to assessing multiple high-level drug-drug and drug-disease adverse drug effects. *Expert Opinion on Drug Metabolism & Toxicology*, 20(7), 579-592. <https://doi.org/10.1080/17425255.2023.2299337>
- Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13), i457-i466. <https://doi.org/10.1093/bioinformatics/bty294>