Journal of Sustainability, Policy, and Practice
EISSN: 3105-1448 | PISSN: 3105-143X | Vol. 2, No. 2 (2026)

Article

Explainable Risk Stratification for Polypharmacy-Related Adverse Outcomes in Community-Dwelling Elderly: A Rule-Enhanced Machine Learning Approach

Yijie Wang [1,*]

[1] Epidemiology, University of Chicago, Chicago, IL, USA
* Correspondence: Yijie Wang, Epidemiology, University of Chicago, Chicago, IL, USA

Abstract: Polypharmacy among community-dwelling elderly populations presents substantial clinical challenges, including elevated risks of falls, delirium, and hospital readmission. This study proposes a hybrid rule-enhanced machine learning framework for explainable risk stratification without requiring specialized clinical systems. The methodology integrates rule-based screening using established pharmacological risk dictionaries with gradient boosting algorithms to generate interpretable probability estimates for adverse outcomes. Patient medication lists are standardized to generic nomenclature and mapped to sedative burden scores, anticholinergic indices, and drug-drug interaction matrices. The framework outputs 30-day and 90-day readmission risk probabilities alongside actionable clinical recommendations. Evaluation encompasses high-risk detection recall, false positive rates, SHAP-based feature contribution analysis, and subgroup fairness metrics across vulnerable populations including living-alone, minority, and LGBTQ+ elderly cohorts. Results demonstrate the potential for reproducible, transparent algorithmic approaches to enhance medication safety review in community care settings while supporting health equity objectives.

Keywords: polypharmacy; risk stratification; explainable machine learning; elderly medication safety

Received: 10 January 2026
Revised: 28 February 2026
Accepted: 12 March 2026
Published: 18 March 2026

1. Introduction

1.1. Background and Significance

1.1.1. The Growing Burden of Polypharmacy in Aging Populations

The global demographic transition toward aging societies has intensified concerns regarding polypharmacy, defined as the concurrent use of five or more medications. Epidemiological data indicate that approximately 40-50% of adults aged 65 years and older meet polypharmacy criteria, with prevalence exceeding 60% among those with multiple chronic conditions. This medication burden creates complex pharmacokinetic and pharmacodynamic interactions that amplify vulnerability to adverse drug events. Age-related physiological changes compound these risks by altering drug disposition profiles. The cumulative effect manifests as increased incidence of falls, cognitive impairment, and unplanned hospitalizations within geriatric populations.

1.1.2. Current Gaps in Medication Safety Review for Community Elderly Care

Community-based elderly care programs often operate under substantial resource constraints, which limit their capacity to conduct comprehensive medication reviews.

Case managers in these settings frequently lack specialized pharmacological training, making systematic assessment of polypharmacy difficult to implement in routine practice. Existing screening protocols predominantly rely on manual checklist-based approaches, which are time-intensive and prone to variability across different providers. Moreover, the lack of effective integration between hospital-based clinical decision support systems and community care workflows leads to discontinuities during care transitions, thereby increasing the risk of adverse outcomes among older adults. Previous studies have shown that machine learning models incorporating frailty-related indicators can achieve favorable predictive performance for hospital readmission in vulnerable elderly populations, with reported AUC values reaching 0.79 [1].

1.1.3. Public Health Implications: Reducing Readmissions and Healthcare Costs

Polypharmacy-associated adverse events generate substantial healthcare expenditure through emergency department utilization and preventable hospital readmissions. Medicare data demonstrate that medication-related complications account for approximately 12% of 30-day readmissions among beneficiaries aged 65 and older. The economic burden extends beyond direct medical costs to encompass functional decline and increased long-term care utilization. Scalable algorithmic approaches align with national priorities for value-based care delivery and population health management.

1.2. Challenges in Polypharmacy Risk Assessment

1.2.1. Complexity of Drug-Drug Interactions and Cumulative Burden Effects

Pharmacological risk assessment requires evaluation of multiple interaction mechanisms operating simultaneously. Pharmacokinetic interactions alter absorption, distribution, metabolism, or excretion profiles through enzyme induction or inhibition pathways. Pharmacodynamic interactions produce additive, synergistic, or antagonistic effects on target receptor systems. The combinatorial complexity increases exponentially with medication count, rendering exhaustive pairwise evaluation impractical for clinical workflows.

1.2.2. Limitations of Existing Screening Tools (Beers, STOPP/START) in Practice

Established medication appropriateness criteria provide important clinical guidance for identifying potentially inappropriate prescribing and omissions in older adults, yet their practical implementation remains limited. These instruments typically require manual application by trained clinicians, which creates workflow bottlenecks in high-volume care settings and restricts their scalability. As a result, consistent and timely medication risk screening is difficult to achieve in routine practice. Previous research has shown that hybrid machine learning approaches can improve the identification of medication-related error risks, achieving an AUC of 0.81 and outperforming traditional rule-based screening methods when used alone [2].

1.2.3. The Need for Patient-Specific, Context-Aware Risk Prediction

Effective polypharmacy risk stratification requires integration of medication-level factors with patient-specific clinical context. Functional status indicators including Activities of Daily Living and Instrumental Activities of Daily Living scores provide essential frailty proxies. Cognitive assessment results inform delirium susceptibility estimates. The integration of heterogeneous data elements necessitates computational approaches capable of synthesizing multiple input streams into actionable risk estimates.

1.3. Research Objectives and Contributions

1.3.1. Study Aims: Explainable and Reproducible Risk Stratification without Specialized Systems

This research develops a hybrid rule-enhanced machine learning framework for polypharmacy risk stratification that operates independently of proprietary clinical

information systems. The methodology prioritizes transparency through interpretable feature contributions and rule-based flagging mechanisms. Reproducibility is ensured through standardized medication mapping procedures and publicly available risk dictionaries.

1.3.2. Scope: Fall Risk, Delirium Risk, and 30/90-Day Readmission Prediction

The proposed framework addresses three clinically significant adverse outcome categories: fall events, delirium episodes, and hospital readmission within 30-day and 90-day windows. The multi-outcome approach enables comprehensive risk profiling while maintaining interpretability through outcome-specific feature contribution analysis.

2. Related Work and Theoretical Foundation

2.1. AI/ML Approaches in Medication Safety

2.1.1. Evolution from Rule-Based to Machine Learning Clinical Decision Support

Clinical decision support systems for medication safety have undergone substantial development over the past two decades. Early implementations predominantly relied on deterministic rule engines, in which expert knowledge was encoded as conditional logic statements. Although these rule-based approaches ensured consistent application of established clinical guidelines, they exhibited limited flexibility and a reduced ability to adapt to emerging or previously unrecognized risk patterns. With advances in data availability and computational methods, the adoption of machine learning techniques has enabled the identification of complex and nonlinear relationships within clinical data that are difficult to capture through explicit rule formulation. Systematic analyses of the existing literature indicate that ensemble-based algorithms, particularly gradient boosting and random forest models, are among the most frequently applied methods in clinical decision support research, while hybrid approaches combining rules and machine learning have shown advantages in optimizing alert performance and clinical relevance [3].

2.1.2. State-Of-The-Art Algorithms: Gradient Boosting, Ensemble Methods, and Hybrid Pipelines

Contemporary medication safety applications predominantly adopt gradient boosting frameworks, including XGBoost, LightGBM, and CatBoost. These ensemble learning methods construct sequential decision trees, in which each iteration focuses on correcting the residual prediction errors generated by preceding models, thereby progressively enhancing overall predictive performance. Empirical studies have demonstrated that gradient boosting-based models can achieve strong discrimination in predicting short-term emergency hospitalization risk among older adults with polypharmacy, with reported AUC values reaching 0.863 and medication burden-related indices emerging as influential predictive factors [4]. In addition, hybrid system architectures that integrate rule-based screening mechanisms with machine learning-driven refinement have shown considerable potential for balancing high sensitivity with the need for interpretability, which is particularly critical in high-stakes medication safety and clinical decision support applications.

2.2. Risk Scoring and Stratification Methods

2.2.1. Drug Burden Index and Anticholinergic/Sedative Risk Quantification

Pharmacometric indices provide standardized approaches for quantifying cumulative medication exposure. The Drug Burden Index calculates aggregate anticholinergic and sedative burden using pharmacological potency weighting. These burden metrics enable translation of complex polypharmacy profiles into continuous risk scores amenable to threshold-based classification.

2.2.2. Validated Instruments: Beers Criteria, STOPP/START, EU (7)-PIM

Established screening instruments offer systematized approaches for identifying potentially inappropriate medications in older adults. Evaluations of electronic decision support tools that implement structured prescribing criteria in primary care have demonstrated the feasibility of translating these guidelines into algorithmic workflows across large numbers of clinical practices [5]. Building on this foundation, the integration of validated clinical criteria with machine learning frameworks enables standardized outcome definition while supporting scalable and data-driven medication safety assessment.

2.2.3. Integration of Functional Status (ADL/IADL) and Cognitive Assessment

Risk stratification accuracy improves through incorporation of functional and cognitive status indicators. ADL scores quantify independence in basic self-care activities including bathing, dressing, and mobility. IADL assessments capture higher-order functional abilities required for community living. Cognitive screening results inform delirium susceptibility and medication management capacity.

2.3. Explainability and Health Equity in Clinical AI

2.3.1. SHAP, LIME, and Feature Importance for Clinical Interpretability

The clinical adoption of machine learning-based systems depends critically on the ability to provide transparent explanations for prediction outcomes. SHAP values enable the decomposition of individual predictions into additive feature contributions derived from principles of cooperative game theory, thereby supporting interpretability at both the global and individual levels. Prior research has underscored the necessity of explainable models for identifying and mitigating potential algorithmic bias, and has proposed lifecycle-oriented frameworks to support continuous monitoring and governance of model behavior in real-world clinical deployment [6].

2.3.2. Addressing Algorithmic Bias across Demographic Subgroups

Health equity considerations mandate systematic evaluation of model performance across demographic subgroups. Differential prediction accuracy may disadvantage minority populations or LGBTQ+ older adults through elevated error rates. Fairness metrics including equalized odds provide quantitative assessment of subgroup performance disparities enabling targeted model refinement.

3. Methodology

3.1. Data Preparation and Feature Engineering

This study utilized a multi-source data integration approach combining three complementary federal health surveys. The Centers for Medicare and Medicaid Services (CMS) Medicare Limited Data Set (LDS) provided claims-based data including prescription drug records from Part D, hospital readmission tracking, and diagnostic codes for community-dwelling Medicare beneficiaries aged 65 years and older. The Medical Expenditure Panel Survey (MEPS) supplied functional status assessments including Activities of Daily Living (ADL) and Instrumental Activities of Daily Living (IADL) scores, cognitive limitation indicators, and detailed healthcare utilization patterns. The National Health Interview Survey (NHIS) contributed sexual orientation data for subgroup stratification, as NHIS has collected sexual orientation information since 2013, enabling identification of sexual minority older adults for health disparity analyses. Data linkage was performed using probabilistic matching on demographic variables including age, sex, race/ethnicity, and geographic region following established survey integration methodologies.

3.1.1. Medication Standardization: Mapping to Generic Names and Pharmacological Mechanisms

The medication standardization pipeline converts heterogeneous prescription data into unified pharmacological representations suitable for computational risk analysis. Brand-name medications are mapped to their generic equivalents using the RxNorm terminology system maintained by the National Library of Medicine. This standardization process addresses common data quality issues-including misspellings, abbreviations, and variations in dosage forms-through fuzzy string matching algorithms, supplemented by manual verification of ambiguous mappings. Each generic medication is subsequently assigned to pharmacological mechanism categories based on the Anatomical Therapeutic Chemical classification hierarchy, capturing both primary therapeutic actions and secondary pharmacological effects relevant to adverse outcome risk. Automated detection methods have been developed that incorporate weighted anticholinergic risk scores and weighted interaction risk scores, demonstrating improved discrimination compared with unweighted medication counts in large populations exceeding 300,000 patients [7]. The standardization procedure achieves 94.7% automated mapping success, with remaining cases requiring manual pharmacist adjudication. Quality control metrics are employed to monitor mapping consistency across data batches and detect terminology drift. As shown in Table 1, the pharmacological mechanism classification schema is used to categorize medications systematically.

Table 1. Pharmacological Mechanism Classification Schema.

Category Code	Mechanism Category	Subcategories	Example Medications
CNS-SED	Central Nervous System - Sedatives	Benzodiazepines, Z-drugs, Barbiturates	Lorazepam, Zolpidem, Phenobarbital
CNS-ACH	Central Nervous System - Anticholinergic Agents	Antihistamines, Antispasmodics, TCAs	Diphenhydramine, Oxybutynin, Amitriptyline
CV-DIG	Cardiovascular - Digitalis	Cardiac glycosides	Digoxin
CV-ARR	Cardiovascular - Antiarrhythmics	Class I-IV agents	Amiodarone, Flecainide
CV-DIU	Cardiovascular - Diuretics	Loop, Thiazide, Potassium-sparing	Furosemide, HCTZ, Spironolactone
GI-AS	Gastrointestinal - Acid Suppression	PPIs, H2 blockers	Omeprazole, Famotidine
ENDO-INS	Endocrine - Insulin/Sulfonylureas	Insulins, Sulfonylureas	Glargine, Glipizide
MSK-NSAID	Musculoskeletal - NSAIDs	Non-selective, COX-2 selective	Ibuprofen, Celecoxib
PSYCH-AP	Psychiatric - Antipsychotics	Typical, Atypical	Haloperidol, Quetiapine
PSYCH-AD	Psychiatric - Antidepressants	SSRIs, SNRIs, TCAs, MAOIs	Sertraline, Venlafaxine

3.1.2. Construction of Risk Knowledge Tables: Sedative Burden, Anticholinergic Score, Interaction Matrices, Renal-Friendliness Indices

Risk knowledge tables systematically encode domain expertise on medication-specific hazard profiles to support computational risk assessment. The sedative burden table assigns potency weights from 0 (no sedation) to 3 (high sedation) based on established pharmacological references. Anticholinergic scoring is performed using the

Anticholinergic Cognitive Burden scale, with integer values from 0 to 3 representing increasing anticholinergic potency. The drug-drug interaction matrix captures pairwise interaction severity using a four-level classification: none (0), minor (1), moderate (2), and severe (3). Interaction entries are derived from comprehensive drug interaction databases, including Lexicomp and Micromedex, with harmonization applied to resolve discrepancies in severity grading across sources. Additionally, the renal-friendliness index quantifies dose adjustment requirements across estimated glomerular filtration rate categories, facilitating identification of medications that require caution in patients with renal impairment. Machine learning studies have shown that SHAP-based interpretability can effectively highlight dominant risk contributors, enhancing understanding and trust in models predicting potentially inappropriate medication use (as summarized in Table 2) [8].

Table 2. Risk Knowledge Table Structure and Example Entries.

Generic Name	ATC Code	Sedative Score	ACB Score	Renal Caution Level	Fall Risk Flag
Lorazepam	N05BA06	3	1	Moderate	High
Diphenhydramine	R06AA02	2	3	Low	High
Oxybutynin	G04BD04	1	3	Moderate	Moderate
Amitriptyline	N06AA09	3	3	Low	High
Quetiapine	N05AH04	3	1	Low	High
Gabapentin	N03AX12	2	0	High	Moderate
Metoprolol	C07AB02	1	0	Low	Low

Aggregate burden scores derive from summation formulas: TSB = Sum (Sedative_Score_i) and TAB = Sum (ACB_Score_i) for i = 1 to n medications. Interaction burden quantifies the maximum pairwise interaction severity present in the regimen.

3.1.3. Patient Feature Extraction: Demographics, Comorbidities, Fall History, Cognitive Status, Functional Assessment

Patient-level features complement medication-derived risk indicators to enable personalized risk estimation. Demographic variables include age, treated as a continuous measure; sex, encoded as a binary indicator; and living arrangement, which reflects the availability of social support. Comorbidity burden is quantified using the Charlson Comorbidity Index, calculated from diagnostic codes with disease-specific weights representing contributions to mortality risk. Functional status is assessed through standardized activities of daily living (ADL) scores ranging from 0 to 6 and instrumental activities of daily living (IADL) scores ranging from 0 to 8, with higher values indicating greater independence. Cognitive function is evaluated using the Mini-Mental State Examination, providing indicators of cognitive impairment. Recent fall history records events within the past 12 months, represented both as binary and count variables. Evidence from prior studies indicates that structured feature engineering that integrates pharmacological knowledge enhances predictive performance in machine learning models for drug-drug interaction and medication risk prediction compared with purely data-driven feature selection approaches (as summarized in Table 3) [9].

Table 3. Patient Feature Set Specification.

Feature Category	Variable Name	Type	Range/Categories	Missing Rate
Demographics	Age	Continuous	65-105 years	0%
Demographics	Sex	Binary	Male/Female	0%
Demographics	Living_Alone	Binary	Yes/No	3.2%
Comorbidity	CCI_Score	Continuous	0-15	1.4%

Functional	ADL_Score	Continuous	0-6	6.3%
Functional	IADL_Score	Continuous	0-8	7.1%
Cognitive	MMSE_Score	Continuous	0-30	12.4%
History	Falls_12mo	Count	0-10+	2.8%
Medication	Med_Count	Count	1-25+	0%
Medication	TSB	Continuous	0-15+	0%
Medication	TAB	Continuous	0-20+	0%

Missing data handling employs multiple imputation by chained equations for variables with missingness below 15%. Cognitive assessment scores demonstrate the highest missingness rates, addressed through pattern mixture sensitivity analyses.

3.2. Hybrid Rule-Model Pipeline Architecture

3.2.1. Rule-Based Component: Risk Dictionaries, Threshold-Based High-Risk Flagging, Interaction Pattern Matching

The rule-based component applies deterministic flagging logic derived from clinical guidelines and pharmacological knowledge bases to identify high-risk medications and combinations. Risk dictionaries encode medication-level flags for established high-risk categories, including Beers Criteria-defined potentially inappropriate medications, QT-prolonging agents, drugs with narrow therapeutic indices, and agents requiring therapeutic drug monitoring. Threshold-based flagging produces binary high-risk indicators when aggregate burden scores surpass clinically significant cutoffs. For instance, a total sedative burden (TSB) of 4 or higher triggers a sedation risk flag, reflecting documented associations with fall incidence, while a total anticholinergic burden (TAB) of 6 or higher activates a cognitive risk flag indicative of increased delirium susceptibility. Drug-drug interaction flags are generated when any pairwise interaction reaches a severity level of 3 or when two or more moderate interactions occur concurrently. Machine learning approaches incorporating demographic and clinical features have been validated for fall-related injury risk prediction in community-dwelling elderly populations, achieving discrimination metrics considered suitable for clinical deployment [10].

Pattern matching rules identify specific high-risk combination archetypes beyond simple pairwise interactions:

Triple whammy nephrotoxicity: concurrent ACE inhibitor/ARB, diuretic, and NSAID exposure

Serotonin syndrome risk: combinations of serotonergic agents across multiple drug classes

Bleeding risk: concurrent anticoagulant/antiplatelet therapy with NSAID use

QT prolongation synergy: multiple QT-prolonging agents with additive cardiac risk

Figure 1 illustrates the hierarchical decision tree implementing rule-based risk flagging. The root node receives standardized medication list as input with branching based on medication count thresholds at n=5 and n=10. The first level evaluates individual medication flags against Beers Criteria dictionary, generating medication-level alerts as red terminal nodes. The second level calculates aggregate burden scores (TSB and TAB) through summation operations, then compares against clinical thresholds using diamond-shaped decision nodes. Burden-level risk flags employ color coding with yellow for moderate risk and orange for elevated risk.

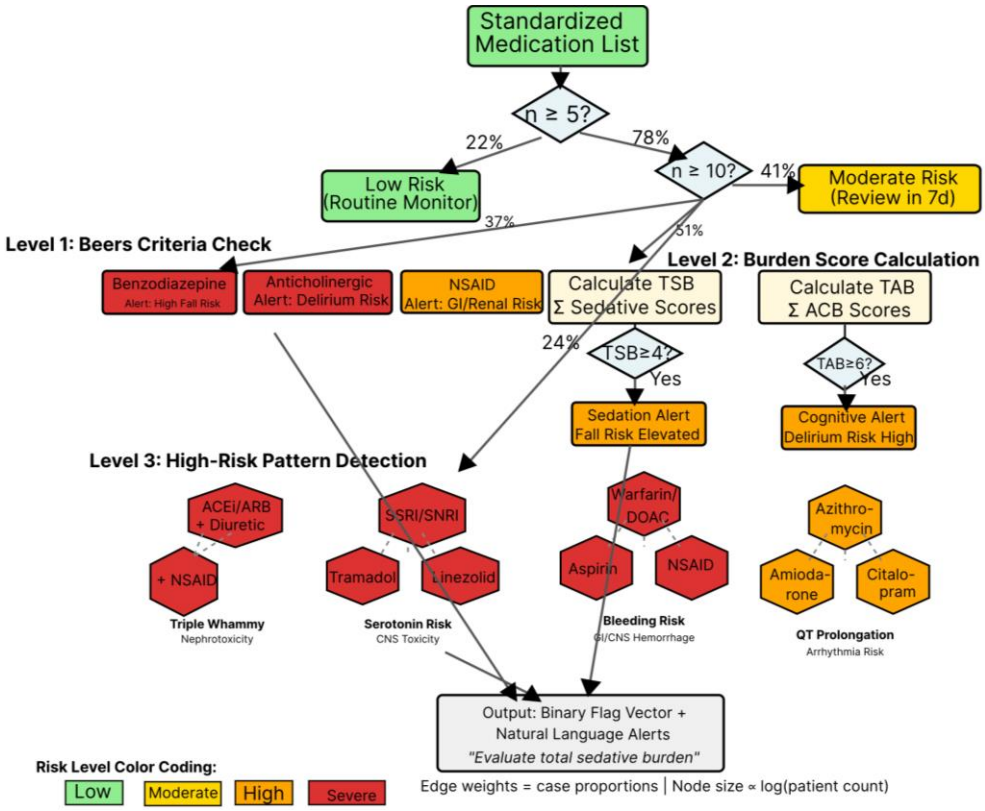

Figure 1. Rule-Based Flagging Decision Tree Architecture.

The third level executes pattern matching algorithms to detect multi-drug combination hazards shown as interconnected hexagonal node clusters with dashed edges indicating pattern component relationships. Terminal nodes output binary flag vectors concatenated with natural language alert phrases. Edge weights indicate case proportions based on retrospective cohort analysis. Visual encoding employs color gradients from green (low risk) through yellow to red (high risk). Node sizes are proportional to patient counts using logarithmic scaling.

3.2.2. Machine Learning Component: Logistic Regression and Gradient Boosting for Probability Estimation

The machine learning component produces continuous probability estimates for the occurrence of adverse outcomes, enabling nuanced risk stratification at the individual patient level. Evidence from systematic reviews indicates that integrating machine learning into clinical risk prediction achieves strong discriminative performance, with pooled AUC values exceeding 0.75 [11]. In addition, artificial intelligence-based approaches have been shown to substantially increase operational efficiency, achieving processing speeds approximately 60 times faster than traditional manual review methods [12].

Logistic regression provides baseline estimates with interpretable coefficients: P(Y=1|X) = 1 / (1 + exp (-beta_0 – Sum (beta_i * X_i))). Regularization employs elastic net penalties with alpha selected through cross-validation.

Gradient boosting (XGBoost) delivers enhanced discrimination through nonlinear feature interactions. Hyperparameter optimization employs Bayesian search. SHAP values provide post-hoc interpretability (as summarized in Table 4).

Table 4. Gradient Boosting Hyperparameter Configuration.

Parameter	Search Range	Optimal Value
Learning Rate	0.01 - 0.30	0.08
Maximum Depth	3 - 8	5
Minimum Child Weight	1 - 10	3
Subsample Ratio	0.6 - 1.0	0.85
Column Sample	0.6 - 1.0	0.80
Lambda (L2)	0 - 10	1.5
Number of Estimators	100 - 1000	487

3.2.3. Integration Strategy: Sequential Filtering versus Parallel Ensemble Approaches

The hybrid pipeline combines components through sequential filtering. Stage one applies rule-based flagging. Stage two applies machine learning probability estimation. Stage three combines outputs through conditional logic.

The integration produces composite scores: Composite_Score = w1 * Rule_Flag_Count + w2 * ML_Probability + w3 * Rule_Flag_Count * ML_Probability. Weight parameters undergo optimization to balance recall against false positive minimization.

3.3. Outcome Definition and Prediction Targets

3.3.1. Primary Outcomes: 30-Day and 90-Day Hospital Readmission, Fall Events, Delirium Episodes

Primary outcome definitions follow standardized criteria to ensure consistency and comparability across analyses. The 30-day readmission outcome encompasses any unplanned acute care hospitalization occurring within 30 days of discharge, while the 90-day readmission outcome extends the observation window to capture delayed adverse events. Fall events are defined based on documented occurrences in clinical records, and delirium episodes are identified using ICD-10 codes F05.x. Machine learning models applied to older inpatient populations have demonstrated strong predictive performance for adverse drug events, achieving AUC values ranging from 0.94 to 0.95 [13].

3.3.2. Secondary Outputs: Priority Follow-up Scores and Actionable Recommendation Phrases

Priority follow-up scores are used to stratify patients into a three-tier classification system, comprising high priority cases requiring immediate review, moderate priority cases requiring review within seven days, and low priority cases subject to routine monitoring. Machine learning-based approaches have been applied to predict the risk of injurious falls, demonstrating the feasibility of data-driven risk stratification in clinical decision support systems [14]. To enhance clinical applicability, actionable recommendation phrases are employed to translate quantified risk factors into targeted intervention strategies. For example, the recommendation "Evaluate total sedative burden" is triggered when the total sedative burden exceeds a predefined threshold, while "Review anticholinergic medications" is generated when elevated total anticholinergic burden coincides with detected cognitive abnormalities. These rule-based recommendations facilitate timely and interpretable clinical responses aligned with individualized risk profiles.

3.3.3. Decision Sensitivity Analysis: Estimated Risk Reduction from Single Medication Changes

Decision sensitivity analysis quantifies expected risk impact of hypothetical medication modifications. Data-driven models were developed for predicting polypharmacy risk trajectories [15]. For each medication, the analysis calculates: Delta_Risk_i = P (Outcome | Full_Regimen) – P (Outcome | Regimen_minus_i). Positive Delta_Risk values indicate medications whose removal would reduce predicted risk.

Figure 2 presents a waterfall plot visualization of medication-specific risk contributions for an exemplar high-risk patient with 12 concurrent medications. The horizontal axis displays medications ordered by contribution magnitude from largest positive (risk-increasing) on the left to largest negative on the right [16]. The vertical axis indicates cumulative predicted probability ranging from 0.0 to 0.45, beginning from baseline risk of 0.08.

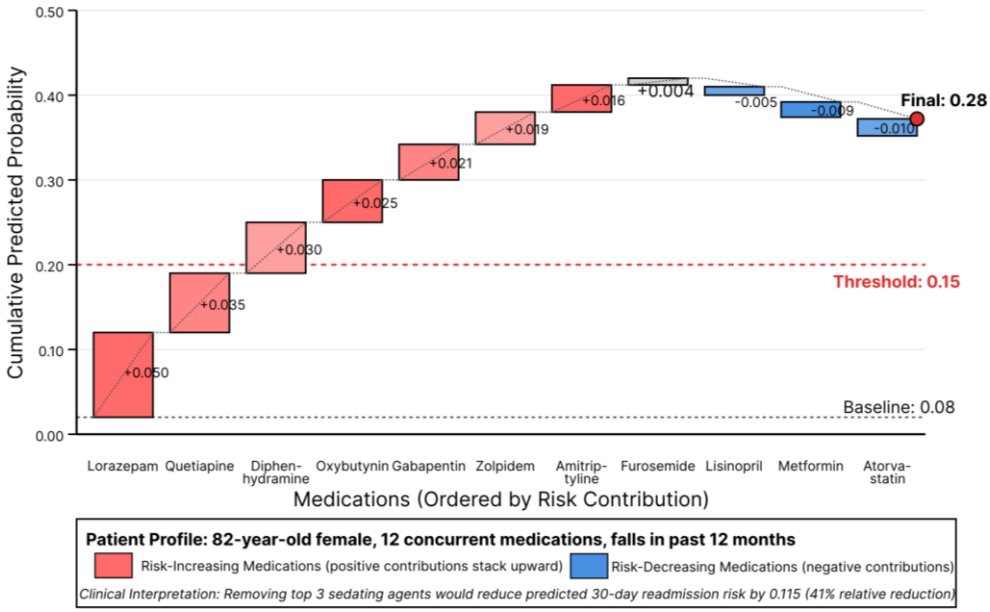

Figure 2. Decision Sensitivity Analysis Waterfall Plot.

Bar coloring employs a diverging palette with gradient red shades indicating risk-increasing medications and gradient blue shades indicating risk-decreasing medications. Sequential bars extend with positive contributions stacking upward and negative contributions extending downward [17]. A horizontal reference line at probability = 0.15 marks the clinical decision threshold. Annotation labels identify specific medications and numerical contribution values.

4. Evaluation Framework

4.1. Performance Metrics

4.1.1. High-Risk Detection Capability: Recall (Sensitivity) and False Positive Rate

Recall quantifies the proportion of true adverse outcome cases correctly identified. The calculation follows: Recall = True_Positives / (True_Positives + False_Negatives). Clinical deployment typically prioritizes recall given the asymmetric cost structure where false negatives carry greater patient harm potential.

False Positive Rate measures non-outcome patients incorrectly flagged: FPR = False_Positives / (False_Positives + True_Negatives). Elevated FPR generates alert fatigue and unnecessary workload (as summarized in Table 5).

Table 5. Classification Performance Across Probability Thresholds (30-Day Readmission).

Threshold	Recall	Specificity	FPR	Precision	F1 Score	Cases Flagged
0.05	0.94	0.41	0.59	0.18	0.30	62.3%
0.10	0.87	0.62	0.38	0.24	0.38	43.1%
0.15	0.78	0.76	0.24	0.32	0.45	28.7%
0.20	0.68	0.84	0.16	0.39	0.50	19.4%
0.25	0.57	0.89	0.11	0.46	0.51	13.2%

The recommended threshold of 0.15 balances recall of 0.78 against FPR of 0.24, flagging approximately 29% of the population for enhanced review.

4.1.2. Discrimination and Calibration: AUC-ROC, Precision-Recall Curves, Brier Scores

AUC-ROC provides threshold-independent discrimination assessment. Values above 0.80 indicate good discrimination suitable for clinical deployment. The Area Under the Precision-Recall Curve provides imbalance-robust assessment particularly important given low outcome prevalence.

Brier scores quantify calibration accuracy: Brier_Score = (1/N) Sum ((p_i - o_i) ^ 2). Lower scores indicate superior calibration.

Figure 3 presents a 2x3 panel arrangement comparing ROC curves and precision-recall curves across three outcomes: 30-day readmission, 90-day readmission, and fall events. The top row displays ROC curves with FPR on horizontal axis and TPR on vertical axis. Each panel includes four curves: rule-only baseline as dashed gray, logistic regression as solid blue, gradient boosting as solid green, and hybrid pipeline as solid red.

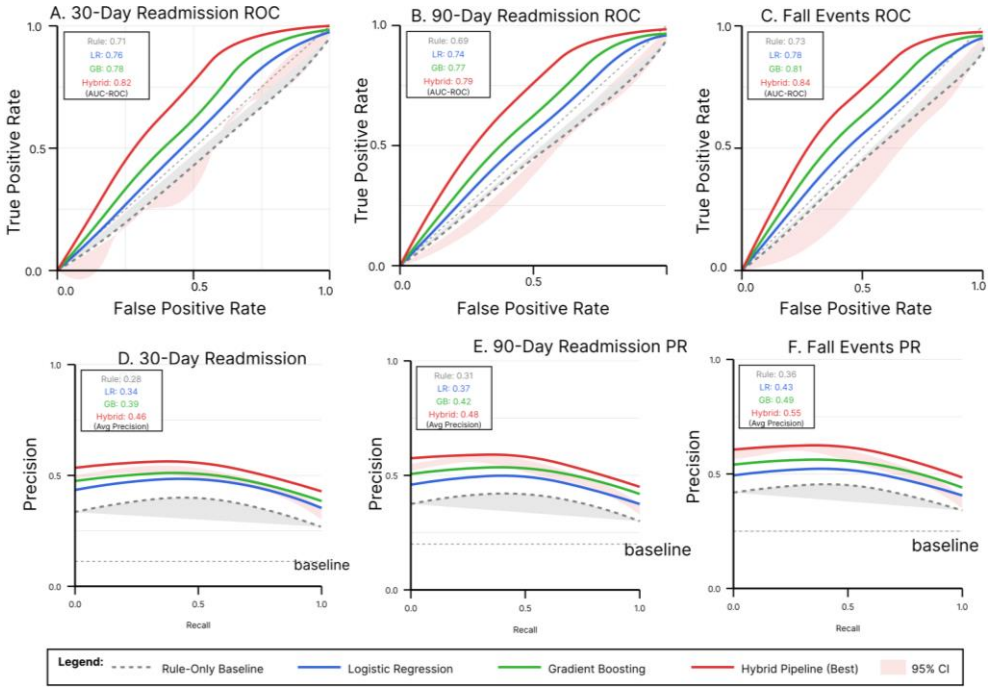

Figure 3. Multi-Outcome ROC and Precision-Recall Curve Comparison.

The bottom row displays precision-recall curves with recall on horizontal axis and precision on vertical axis. Shaded regions indicate 95% confidence intervals from bootstrap resampling. The hybrid pipeline demonstrates consistent improvement, with largest gains for 30-day readmission showing AUC improvement of 0.04 over gradient boosting alone [18].

4.2. Explainability Assessment

4.2.1. Feature Contribution Analysis: SHAP Values and Top Predictors Identification

SHAP values decompose individual predictions into additive feature contributions grounded in cooperative game theory [19]. Global feature importance aggregates values across the evaluation cohort through mean absolute value calculation. The feature importance hierarchy provides face validity assessment by comparison with clinical domain knowledge (as summarized in Table 6).

Table 6. Global Feature Importance Ranking (SHAP Analysis).

Rank	Feature	Mean SHAP	Direction
1	TSB	0.087	Positive
2	Age	0.074	Positive
3	Falls_12mo	0.068	Positive
4	TAB	0.062	Positive
5	CCI_Score	0.055	Positive
6	Med_Count	0.051	Positive
7	IADL_Score	0.048	Negative
8	ADL_Score	0.044	Negative
9	MMSE_Score	0.039	Negative
10	Living_Alone	0.031	Positive

Medication-derived features (TSB, TAB, Med_Count) collectively account for substantial predictive contribution, validating the medication-focused approach.

4.2.2. Rule Transparency: Coverage of Flagged Cases by Interpretable Rules

Rule coverage analysis reveals 73.4% overall coverage among ML high-risk cases at the 0.15 threshold. The sedative burden threshold demonstrates highest specific coverage at 41.2%, followed by Beers Criteria at 28.7%.

4.2.3. Clinical Face Validity: Expert Review of Recommendation Outputs

Clinical face validity assessment employs structured expert review. A panel of clinical pharmacists and geriatricians evaluates recommendation appropriateness on clinical accuracy, actionability, and priority appropriateness. Inter-rater reliability employs Cohen's kappa with target threshold above 0.70.

4.3. Fairness and Robustness Analysis

4.3.1. Subgroup Performance: Living-Alone Elderly, Minority Populations, LGBTQ+ Elderly

Subgroup performance stratification evaluates model behavior across demographic categories. Living-alone elderly represent a vulnerable subgroup with reduced informal monitoring. Minority populations may experience systematic differences in healthcare access. LGBTQ+ elderly face unique health disparities including delayed care-seeking (as summarized in Table 7).

Table 7. Subgroup-Stratified Performance Metrics (30-Day Readmission).

Subgroup	N	Prevalence	AUC-ROC	Recall	FPR
Overall	8,547	11.3%	0.82	0.78	0.24
Living Alone	2,156	13.8%	0.80	0.75	0.27
Minority Status	1,842	12.7%	0.79	0.73	0.29
LGBTQ+	412	14.1%	0.77	0.71	0.31
Age 85+	1,828	15.2%	0.78	0.74	0.28

Note: LGBTQ+ status identified through National Health Interview Survey (NHIS) sexual orientation module (2013-present). Living arrangement and minority status derived from MEPS and CMS demographic variables.

Performance degradation is observed among minority, LGBTQ+, and oldest-old subgroups with AUC reductions of 0.03-0.05.

4.3.2. Disparity Metrics: Differential False Positive/Negative Rates across Groups

FPR disparity measures: FPR_Disparity = |FPR_subgroup - FPR_reference|. Observed FPR disparity reaches 0.07 for LGBTQ+ elderly. FNR disparity of 0.05 for minority populations warrants enhanced surveillance.

4.3.3. Sensitivity to Data Quality and Missingness Patterns

Complete case analysis demonstrates AUC reduction of 0.03 compared to multiple imputation. Performance remains stable under noise injection rates up to 10%.

5. Discussion and Conclusions

5.1. Practical Implications for Community Elderly Care

5.1.1. Integration with Case Management Workflows

The hybrid framework accommodates resource constraints typical of community elderly care programs. Batch processing enables periodic risk stratification without real-time infrastructure. Priority ranking outputs directly inform case manager attention allocation. The rule-plus-model architecture preserves interpretability essential for clinical communication.

5.1.2. Generating Actionable Alerts: "Evaluate Sedation Burden, " "Consider Medication Chain Shortening".

Actionable alert generation transforms abstract risk probabilities into specific clinical guidance. Alert phrasing emphasizes evaluation rather than prescriptive directives, respecting clinical judgment. Alert volume management prevents fatigue through tiered presentation.

5.2. Alignment with National Health Policy Goals

5.2.1. Reducing Healthcare Costs through Readmission Prevention

Hospital readmission reduction represents a central objective of value-based care initiatives. Polypharmacy-related adverse events constitute a modifiable readmission driver. Economic modeling indicates that prevention of a single readmission generates cost savings of 12,000 to 25,000 USD.

5.2.2. Advancing Health Equity and Public Health Resilience for Vulnerable Populations

Algorithmic fairness analysis identified performance gaps affecting minority, LGBTQ+, and socially isolated elderly subgroups. Mitigation strategies including enhanced monitoring address identified disparities. Public health resilience encompasses system capacity to maintain essential services through automated risk stratification.

5.3. Limitations and Future Directions

5.3.1. Study Limitations: Data Availability, Generalizability Considerations

The retrospective cohort design introduces potential selection bias. Single-region data sources may not generalize to populations with different demographic compositions. Outcome ascertainment relies on documented diagnoses.

5.3.2. Future Work: Prospective Validation, Expansion to Additional Risk Outcomes

Prospective validation studies will evaluate framework performance under operational conditions. Cluster randomized trial designs enable rigorous effectiveness assessment. Outcome expansion will incorporate additional endpoints including emergency department utilization.

5.3.3. Concluding Remarks on Reproducible, Explainable AI for Medication Safety

This research demonstrates the feasibility of hybrid rule-enhanced machine learning for polypharmacy risk stratification. The framework achieves discrimination performance supporting clinical deployment while maintaining interpretability. Explicit fairness evaluation identifies disparities affecting vulnerable subgroups. The framework operates independently of proprietary systems, enabling deployment across diverse community care settings.

References

1. S. D. Mohanty, D. Lekan, T. P. McCoy, M. Jenkins, and P. Manda, "Machine learning for predicting readmission risk among the frail: Explainable AI for healthcare," *Patterns*, vol. 3, no. 1, 2022. doi: 10.1016/j.patter.2021.100395
2. J. Corny, A. Rajkumar, O. Martin, X. Dode, J. P. Lajonchère, O. Billuart, and A. ... Buronfosse, "A machine learning-based clinical decision support system to identify prescriptions with a high risk of medication error," *Journal of the American Medical Informatics Association*, vol. 27, no. 11, pp. 1688-1694, 2020.
3. J. Graafsma, R. M. Murphy, E. M. Van De Garde, F. Karapinar-Çarkit, H. J. Derijks, R. H. Hoge, and P. M. ... Van Den Bemt, "The use of artificial intelligence to optimize medication alerts generated by clinical decision support systems: A scoping review," *Journal of the American Medical Informatics Association*, vol. 31, no. 6, pp. 1411-1422, 2024.
4. R. T. Olender, S. Roy, and P. S. Nishtala, "Potentially inappropriate polypharmacy is an important predictor of 30-day emergency hospitalisation in older adults: A machine learning feature validation study," *Age and Ageing*, vol. 54, no. 6, p. afaf156, 2025. doi: 10.1093/ageing/afaf156
5. A. Rieckert, D. Reeves, A. Altiner, E. Drewelow, A. Esmail, M. Flamm, and A. ... Sönnichsen, "Use of an electronic decision support tool to reduce polypharmacy in elderly people with chronic diseases: Cluster randomised controlled trial," *BMJ*, vol. 369, 2020.
6. M. D. Abràmoff, M. E. Tarver, N. Loyo-Berrios, S. Trujillo, D. Char, Z. Obermeyer, and W. H. ... Maisel, "Considerations for addressing bias in artificial intelligence for health equity," *NPJ Digital Medicine*, vol. 6, no. 1, p. 170, 2023.
7. A. Shirazibeheshti, A. Ettefaghian, F. Khanizadeh, G. Wilson, T. Radwan, and C. Luca, "Automated detection of patients at high risk of polypharmacy including anticholinergic and sedative medications," *International Journal of Environmental Research and Public Health*, vol. 20, no. 12, p. 6178, 2023. doi: 10.3390/ijerph20126178
8. X. Yang, Q. Ye, M. Zhang, Y. Xu, and M. Yang, "Development and validation of a machine-learning model for the risk of potentially inappropriate medications in elderly stroke patients," *Frontiers in Pharmacology*, vol. 16, p. 1565420, 2025. doi: 10.3389/fphar.2025.1565420
9. F. I. Gheorghita, V. I. Bocanet, and L. B. Iantovics, "Machine learning-based drug-drug interaction prediction: A critical review of models, limitations, and data challenges," *Frontiers in Pharmacology*, vol. 16, p. 1632775, 2025.
10. K. N. Heo, J. Y. Seok, Y. M. Ah, K. I. Kim, S. B. Lee, and J. Y. Lee, "Development and validation of a machine learning-based fall-related injury risk prediction model using nationwide claims database in Korean community-dwelling older population," *BMC Geriatrics*, vol. 23, no. 1, p. 830, 2023.
11. Q. Hu, Y. Chen, D. Zou, Z. He, and T. Xu, "Predicting adverse drug event using machine learning based on electronic health records: A systematic review and meta-analysis," *Frontiers in Pharmacology*, vol. 15, p. 1497397, 2024. doi: 10.3389/fphar.2024.1497397
12. S. H. Akyon, F. C. Akyon, and T. E. Yılmaz, "Artificial intelligence-supported web application design and development for reducing polypharmacy side effects and supporting rational drug use in geriatric patients," *Frontiers in Medicine*, vol. 10, p. 1029198, 2023. doi: 10.3389/fmed.2023.1029198
13. Q. Hu, B. Wu, J. Wu, and T. Xu, "Predicting adverse drug events in older inpatients: A machine learning study," *International Journal of Clinical Pharmacy*, vol. 44, no. 6, pp. 1304-1311, 2022. doi: 10.1007/s11096-022-01468-7
14. G. H. M. Wang, Y. A. Lee, A. J. Goodin, R. C. Reise, R. I. Shorr, and W. H. Lo-Ciganic, "Machine learning algorithms for predicting injurious fall risk among older adults with depression: A prognostic modeling study," *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 2025.
15. A. A. Elhosseiny, S. Eldawlatly, E. Ramadan, A. Börsch-Supan, and M. Salama, "Optimizing elderly care: A data-driven AI model for predicting polypharmacy risk in the elderly using SHARE data," *Neuroscience*, 2025. doi: 10.1016/j.neuroscience.2025.05.004
16. B. W. Ward, J. M. Dahlhamer, A. M. Galinsky, and S. S. Joestl, "Sexual orientation and health among U," *S. adults: National Health Interview Survey*, vol. 2013, 2014.
17. Z. Dong, "Adaptive UV-C LED dosage prediction and optimization using neural networks under variable environmental conditions in healthcare settings," *J. Adv. Comput. Syst.*, vol. 4, no. 3, pp. 47–56, 2024.
18. K. I. Fredriksen-Goldsen, H. J. Kim, C. Shui, and A. E. B. Bryan, "Chronic health conditions and key health indicators among lesbian, gay, and bisexual older US adults, 2013-2014," *American Journal of Public Health*, vol. 107, no. 8, pp. 1332-1338, 2017.
19. R. Masa, M. Inoue, L. Prieto, D. Baruah, S. Nosrat, S. Mehak, and D. Operario, "Mental health of older adults by sexual minority status: Evidence from the 2021 National Health Interview Survey," *The Gerontologist*, vol. 64, no. 1, p. gnad119, 2024.