International Journal of Sustainability and Innovation in Engineering (IJSIE)
https://www.doi.org/10.56830/IJSIE2026
Vol. 4 2026

A Governed Machine Learning Framework for Predicting Healthcare Provider Treatment Adoption and Prescribing Behavior

https://www.doi.org/10.56830/IJSIE202610

Venkat Navneeth Burla
Senior Data Scientist, MA, USA
Email: bnavneetreddy@gmail.com

Received: 17th May, 2026, Accepted: 14th August, 2026, Published: 7th October, 2026

Abstract

Machine learning has not yet been studied in relation to treatment adoption among healthcare providers, even though it is relevant for demand forecasting, evidence prediction, and healthcare resource planning. The researchers developed a governed machine-learning framework to estimate observable use of semaglutide and tirzepatide and link it to prescribing activity among Medicare Part D prescribers. This study used a retrospective longitudinal design, drawing on public CMS data from 2021 to 2024 and a stratified sample of 120,000 prescribers. The study compared regularized logistic regression, random forest, and XGBoost across temporal holdout validation, calibration, SHAP explainability, subgroup assessment, and governance controls. In the provisional 2024 holdout analysis, 3,129 of 16,487 eligible providers crossed the observable-adoption threshold of at least 11 claims, yielding an adoption rate of 19.0%. In the provisional classification analysis, XGBoost achieved the highest AUROC (0.858), AUPRC (0.617), Brier score (0.112), calibration slope (0.98), and approximate F1 score (0.60). In the provisional regression analysis, XGBoost achieved a mean absolute error of 38.6 standardized 30-day fills and an R^2 of 0.61 for post-adoption prescribing intensity. The framework illustrates how provider-behavior prediction can be coupled with provenance, fairness, explainability, provider monitoring, and human oversight. Furthermore, future studies using prospective Medicaid, multi-insurance, and multi-institutional EHR data across reimbursement systems and clinical practice would support external validation before operational deployment.

Keywords: Governed Machine Learning, Healthcare Provider Treatment Adoption, Prescribing Behavior Prediction, Medicare Part D, Explainable Artificial Intelligence, Semaglutide and Tirzepatide Adoption.

1. Introduction

Prescribing behavior is one measurable aspect of the diffusion of treatment since the effectiveness of therapies from regulatory approval to guideline recommendations to actual implementation in Medicare policies depends on prescribing behavior. These decisions affect pharmaceutical needs, payer spending, inventory management, and the rate of evidence incorporation in the health system. Prescription drug spending in the U.S. was $467.0 billion in 2024 and was growing by 7.9% annually. The rapid uptake of incretin therapies shows that provider behavior matters. Data from national retail pharmacy chains revealed that monthly

semaglutide fills grew 442% from 471,876 in January 2021 to 2,555,308 in December 2023. In 2023, Medicare Part D paid for 28.5% of Ozempic fills and 32.9% of Rybelsus fills. This growth can lead to variations in prescription volume by location and specialty. Thus, projecting patient usage alone does not tell researchers which clinicians will accept a therapy. Provider-level prediction can support forecasting, educational planning, and implementation monitoring.

While healthcare machine-learning applications have grown, most focus on predicting whether a patient will develop a diagnosis, experience clinical deterioration, have an adverse outcome, or receive treatment rather than on providers' longitudinal adoption. Existing models also define a good prescription narrowly, focusing more on discrimination metrics than on probability calibration, explainability, subgroup fairness, and lifecycle monitoring. This restriction is operationally significant, as specialties and geographic characteristics, as well as utilization and differences in the mix of beneficiaries represented in provider datasets, may also capture structural differences and cause inappropriate profiling if used without controlling for those factors. Hence, trustworthy healthcare AI systems need to emphasize fairness, traceability, robustness, explainability, and monitoring throughout deployment. The framework for prescribing behaviors therefore needs to be augmented with predictive accuracy, data governance, transparent explanation, human oversight, and post-deployment monitoring.

The research has four objectives. It aims to characterize longer-term trends in supply-side treatment adoption and prescription writing by Medicare Part D providers from 2021 to 2024. It also aims to build and evaluate machine-learning models to predict the following year's uptake of semaglutide- and tirzepatide-containing therapies and subsequent prescribing intensity. The study aims to use explainable machine learning methods to identify provider, practice, and historical prescribing factors most associated with adoption. It also seeks to create and test a governance framework that considers data provenance, calibration, fairness evaluations, explainability, human oversight, and post-deployment monitoring.

The study included only Medicare Part D prescribers identified in CMS administrative claims in the U.S. for 2021–2024. The analytical population focused on high-volume specialties essential for cardiometabolic prescribing, such as internal medicine, family practice, endocrinology, and cardiology. Because prediction was at the provider level, researchers used semaglutide and tirzepatide as the treatment-adoption case. The model did not make prescriptions for specific patients or judge clinical appropriateness, eligibility for treatment, drug dose, contraindications, or therapeutic benefit for the individual patient.

This study consists of several sections. Section 2 reviews the literature on provider treatment adoption, prescription prediction, incretin diffusion, explainable machine learning, and healthcare AI governance to frame the study's research gap. Section 3 reports the data sources, cohort construction, feature engineering, model development, temporal validation, and evaluation procedures. Section 4 describes the machine-learning infrastructure's explainability, provenance, fairness checks, monitoring, and human oversight. Section 5 presents the key results and compares them with the reviewed literature. Section 6 summarizes the study's implications, limitations, and directions for future validation and deployment.

2. Literature Review

2.1.Healthcare Provider Treatment Adoption and Prescribing Behavior

Clinical, organizational, financial, and geographical conditions affect healthcare providers' treatment adoption. Specialties help build brand recognition for a therapy, and experience with similar classes can reduce uncertainty for newer products (Yaseen & Lahiri, 2023). Patient complexity also influences treatment decisions because clinicians may have several treatment options for patients with diabetes, obesity, cardiovascular disease, and other comorbidities. Reimbursement restrictions, formulary status, prior authorization policies, and patient cost-sharing can constrain medicine use, and the amount of medicines dispensed indicates the extent to which medicines are used (Zhang et al., 2026).

Peer behavior and local practice can, however, speed up or slow down uptake, and the rural environment can influence access to specialists and pharmacy availability. Quality of use differs from uptake. Effectiveness asks whether an eligible patient responds to a therapy, and adoption asks whether and when a therapy becomes part of clinical behavior (Wesolowicz et al., 2025). Thus, adoption should be considered not only as a cross-sectional prescription count but also as a diffusion process over time.

2.2.Machine Learning for Predicting Prescribing Decisions

Some strategies involve training ML algorithms on historical healthcare data to approximate prescribing decisions. A random forest can model nonlinear relationships without assuming a specific functional form, while logistic regression can serve as a baseline for estimating probabilities (Efthimiou et al., 2024). Neural networks can learn complex patterns from large electronic health record datasets, and gradient boosting can improve discrimination by sequentially correcting prediction errors.

Prediction via physician profile models considers physician characteristics. In 2023, a deep learning model to predict OUD, in an EHR study of opioid prescriptions, achieved an AUROC of about 0.94. A physician profile study analyzed 174,806 appointments involving 30 physicians and 53,453 patients. Using a physician's digital profile can reduce the prescription prediction model's RMSE from 4.20 to 0.74, an improvement of about 82.4% (Kondrashov et al., 2026). However, caution is needed when interpreting and extrapolating patient-level findings, physician-level findings, and results from 30 physicians to national-level physician adoption.

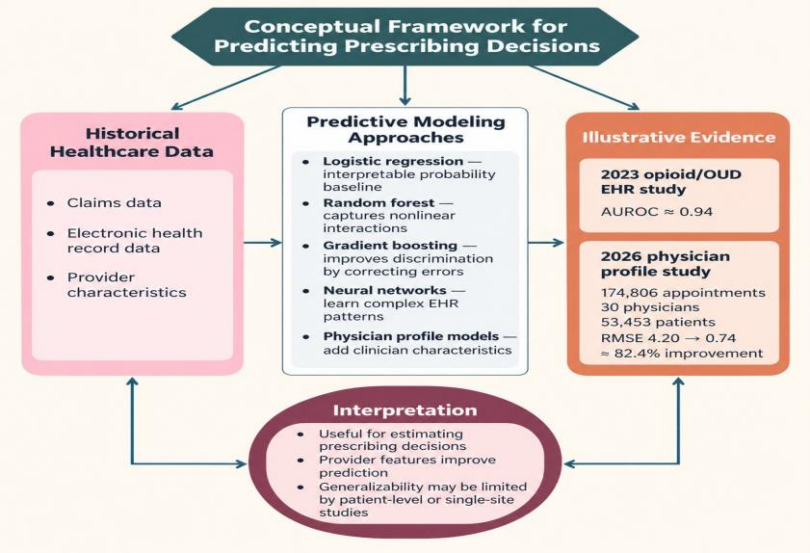


Figure 1: Conceptual framework linking healthcare data, predictive models, empirical evidence, and interpretation of prescribing-decision outcomes.

Figure 1 shows how machine learning methods can be applied in healthcare to predict prescribing decisions. This structure covers the healthcare information spectrum, from historical healthcare data to predictive healthcare data, evidence, and interpretation. It provides an overview of transforming healthcare data into insights using logistic regression, random forests, gradient boosting, neural networks, and the physician profile model, including model applicability, provider attributes and complexities, and limitations on model generalizability.

2.3.Real-World Growth and Access to Incretin Therapies

To understand provider uptake under changing access pathways for semaglutide and tirzepatide, researchers need to examine real-world prescribing trends. Semaglutide prescribing throughout 2021, 2022, and 2023 grew at an accelerated pace, with almost 2.6 million retail fills in December 2023 (Scannell et al., 2024). This was significant because Medicare had become an important payer for incretin therapy. The percentage of Medicare Part D plans covering tirzepatide reportedly increased by 66.4 percentage points, from 26.5% in the third quarter of 2022 to 92.9% in the third quarter of 2024 (Liu et al., 2025). Approximately 83% of the semaglutide and tirzepatide formulations examined required prior authorization during the third quarter of 2024.

The results indicate that factors beyond preference influenced the observed prescribing. Measurable prescribing factors include drug cost, formulary status, beneficiary cost sharing, and coverage limitations, all of which help explain observed prescribing patterns (Kyle et al., 2026). Table 1 summarizes key factors and trends influencing real-world uptake of semaglutide and tirzepatide from 2021 to 2024. The table shows prescription adoption, expanded Medicare Part D coverage, prior authorization requirements, and other non-clinical factors affecting prescription adoption.

Table 1: Real-world growth and access trends influencing incretin therapy adoption between 2021 and 2024.

Indicator	Reported value	Interpretation
Semaglutide monthly retail prescription fills (January 2021–December 2023)	Increased from 471,876 to 2,555,308 fills/month (442% increase)	Demonstrates rapid real-world uptake and increasing clinical demand for incretin therapies
Medicare Part D share of Ozempic fills (2023)	28.5%	Indicates the important role of Medicare beneficiaries in semaglutide utilization
Medicare Part D share of Rybelsus fills (2023)	32.9%	Shows substantial Medicare involvement in oral semaglutide access
Medicare Part D plans covering tirzepatide (Q3 2022)	26.5% of plans	Limited early formulary availability
Medicare Part D plans covering tirzepatide (Q3 2024)	92.9% of plans	Represents a 66.4 percentage-point increase in coverage availability
Semaglutide and tirzepatide formulations requiring prior authorization (Q3 2024)	Approximately 83%	Demonstrates that access restrictions continued despite wider formulary inclusion
Key non-clinical determinants of uptake	Drug cost, formulary inclusion, beneficiary cost sharing, coverage restrictions, administrative burden	Indicates that observed prescribing reflects both provider preference and healthcare access conditions

2.4.Explainability, Fairness, and Healthcare AI Governance

Governance needs to be built into prediction, not prescribed after model development. The FUTURE-AI consortium, made up of 117 experts from 50 countries, organized trustworthy AI for healthcare into six principles of fairness, universality, traceability, usability, robustness, and explainability, and released a set of 30 recommendations for the development and monitoring of AI for healthcare (Lekadir et al., 2025). These principles are especially relevant in settings with historically unequal distribution or specialty-specific applications. Similarly, operational resilience is a key consideration in AI research.

One cited study proposed a fault-tolerant cloud-computing architecture that includes federated learning, deep reinforcement learning, predictive fault detection, dynamic scaling, and self-healing (Lulla & Sayyed, 2025). Although that study concerns mission-critical computing rather than healthcare systems, it presents an engineering paradigm for adding monitoring, failure detection, traceability, and recovery to production systems (Malik & Prashasti, 2023; Johnphill et al., 2026). Provider prediction should also include safeguards against automated assessments without clinical review, opaque commercial targeting, and automated coverage approval.

2.5.Validation and Reporting of Healthcare Prediction Models

To assess healthcare prediction models, a single discrimination statistic is not enough. Temporal validation assesses the model's ability to extrapolate to future periods by training

on the earliest prescribing periods. AUROC measures a model's ability to rank cases correctly, while precision-recall analysis is especially informative when adoption is uncommon. Calibration evaluates agreement between predicted event probabilities and observed event frequencies, and the Brier score measures probabilistic error (Ghazalbash et al., 2026).

Subgroup evaluation results may show variation in performance across specialties, regions, rurality categories, or beneficiary compositions. In TRIPOD+AI (2024), the focus is on clear reporting of the development and testing of the prediction model, while PROBAST+AI (2025) addresses risks with participants and data sources as well as predictors, outcomes, and analysis (Collins et al., 2024; Moons et al., 2025). Cross-domain evidence also supports the multidimensional assessment. Researchers reported an F1 score of 0.88, a 6.2% false-positive rate, and a usability score of 4.3 out of 5 in an AI-enabled infrastructure system (Enugala & Goel, 2025). This evidence supports a multidimensional assessment of discrimination, calibration, reliability, usability, and monitoring.

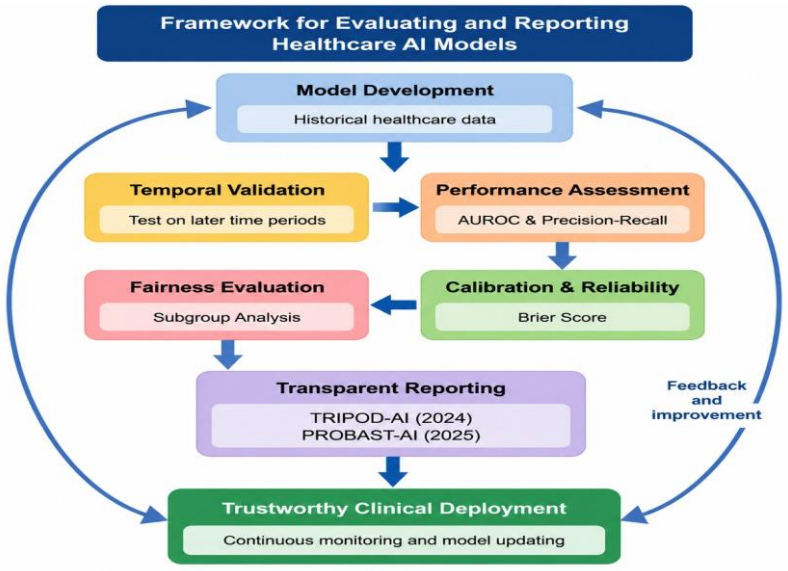


Figure 2: Framework illustrating validation, reporting, and continuous improvement processes for healthcare prediction models.

Figure 2 presents an evaluation and reporting framework for healthcare prediction models, from healthcare dataset extraction to model building, followed by validation, performance assessment, calibration, fairness checking, transparent reporting, and deployment. The overall framework includes AUROC, precision-recall analysis, the Brier score, TRIPOD+AI, and PROBAST+AI as assessment and reporting elements. Feedback mechanisms and continuous monitoring enhance the models' reliability and trustworthiness.

2.6.Limitations and Research Gaps in the Literature

The topic has received some attention, but six major research gaps remain in healthcare prediction. The literature has placed greater emphasis on predicting patient treatment than on modeling provider treatment adoption. Repeated cross-sectional designs are also poorly represented in models, although they are appropriate for studying annual treatment diffusion. Research also lacks evaluations of integrated approaches to calibration, fairness, explainability, and monitoring**.** Moreover, certain proprietary claims and electronic health record (EHR) datasets may be restricted, limiting replication. Public CMS prescription files

also improve data transparency, but limitations still make it difficult to observe and understand low-volume provider use. Additionally, provider-behavior prediction studies often fail to account for governance factors such as provider data provenance, sensitive attributes, provider model risk, human oversight, and lifecycle surveillance.

3. Methods and Techniques

3.1.Data Collection Methods

The study was retrospective and longitudinal, using three Centers for Medicare and Medicaid Services (CMS) public datasets on Medicare Part D prescribing from 2021 to 2024. The first source was Medicare Part D Prescribers by Provider. The 2024 release included 1,416,883 provider records and 84 attributes, including National Provider Identifier, specialty, state, Rural Urban Commuting Area, total claims, standardized 30-day fills, drug cost, low-income subsidy utilization, Medicare Advantage Prescription Drug, stand-alone Prescription Drug Plan claims, and aggregate beneficiary characteristics. The second source was Medicare Part D Prescribers by Provider and Drug for 2021 to 2024. This dataset provided claim counts, NPI, brand and generic drug names, standardized fills, days supplied, and total drug cost for the identification of semaglutide and tirzepatide exposure.

The study measured the outcome as observable treatment adoption rather than absolute first prescribing because CMS suppresses provider-drug observations with 10 or fewer claims; therefore, a published record corresponds to at least 11 claims. The third source was Medicare Part D Spending by Drug 2024, which comes from 100% Final Action Part D prescription claims for aggregate validation of utilization and spending trends. Table 2 summarizes the data sources, study periods, units of analysis, key variables, and analytical purposes used in the study.

Table 2: Data sources, study periods, units of analysis, key variables, and analytical purposes.

Data Source	Study Period	Unit of Analysis	Key Variables Used	Analytical Purpose
Medicare Part D Prescribers by Provider	2021–2024	Individual healthcare prescriber	NPI, specialty, state, RUCA classification, total claims, standardized 30-day fills, drug cost, beneficiary count, brand/generic utilization, MAPD/PDP claims, low-income subsidy utilization, aggregate beneficiary characteristics	Characterization of provider prescribing patterns and development of provider-level prediction features
Medicare Part D Prescribers by Provider and Drug	2021–2024	Provider–drug combination	NPI, brand name, generic name, claim counts, standardized fills, days supplied, total drug cost	Identification of observable semaglutide and tirzepatide exposure and treatment adoption outcomes

Data Source	Study Period	Unit of Analysis	Key Variables Used	Analytical Purpose
Medicare Part D Spending by Drug	2024	Drug-level aggregate prescription records	Final-action Part D prescription claims, utilization measures, spending information	External validation of aggregate drug utilization and expenditure trends
Analytical Cohort	2021–2024 longitudinal observation window	Individual prescriber	Eligibility status, specialty, geographic classification, baseline prescribing activity	Construction of the predictive modeling cohort (N = 120,000 prescribers) using stratified random sampling across specialty, Census region, and urban–rural categories

The study used a sample of 120,000 individual Medicare Part D prescribers. To be eligible, a provider had to appear in the annual provider files for the observation period, practice in the United States or the District of Columbia, have an individual provider designation, have at least 100 Part D claims during the baseline year, and have specialty and geographic information. The study obtained the representation by specialty, Census region, and urban–rural status through stratified random sampling (Gidali & Ward, 2023). This sampling design captured national heterogeneity while minimizing computational demands. The study emphasized security, data control, and traceable processing because healthcare analytics platforms require resilient information-retrieval processes (Sardana et al., 2025).

3.2.Data Analysis

Researchers first harmonized the annual files before linking them. NPI was used only as an identifier and was not interpreted as a numeric feature. Researchers checked provider-year NPI records for duplicates before merging. Structurally unavailable observations were not set to zero; they were retained as suppressed observations. Highly skewed monetary variables were retained in their original form when appropriate or log-transformed, while categorical variables were standardized across release years.

The main outcome identified providers who had no published semaglutide or tirzepatide record in the previous observation year and then had a published record in the following year, corresponding to at least 11 claims (Sargent et al., 2024). Observable adoption therefore represented crossing the CMS public-reporting threshold and did not necessarily indicate a provider's first prescription. A secondary outcome was prescribing intensity, defined as total standardized 30-day fills after adoption. Researchers summarized skewed continuous variables using the median and interquartile range and categorical variables using counts and percentages.

3.3.Machine Learning Model Development and Validation

Researchers applied and compared three prediction models. The researchers used regularized logistic regression as an interpretable baseline, random forest as an example of bagged nonlinear learning, and XGBoost as an example of gradient-boosted classification. The researchers divided the data into 96,000 prescribers for developing the models and 24,000 prescribers for an independent holdout. The researchers used earlier-year provider adoption

transitions to train and tune the models and used 2023-to-2024 provider adoption transitions as the primary temporal evaluation sample, as eligibility allowed.

The study did not include NPI, provider name, or exact address as predictors. Researchers used grouped cross-validation to avoid leaking fold information from the same provider between training and validation (Fourkiotis & Tsadiras, 2025). Classification performance was assessed by AUROC, AUPRC, precision, recall, F1 score, Brier score, calibration slope, and calibration intercept. The researchers used mean absolute error, root mean squared error, and R^2 to evaluate post-adoption prescribing intensity.

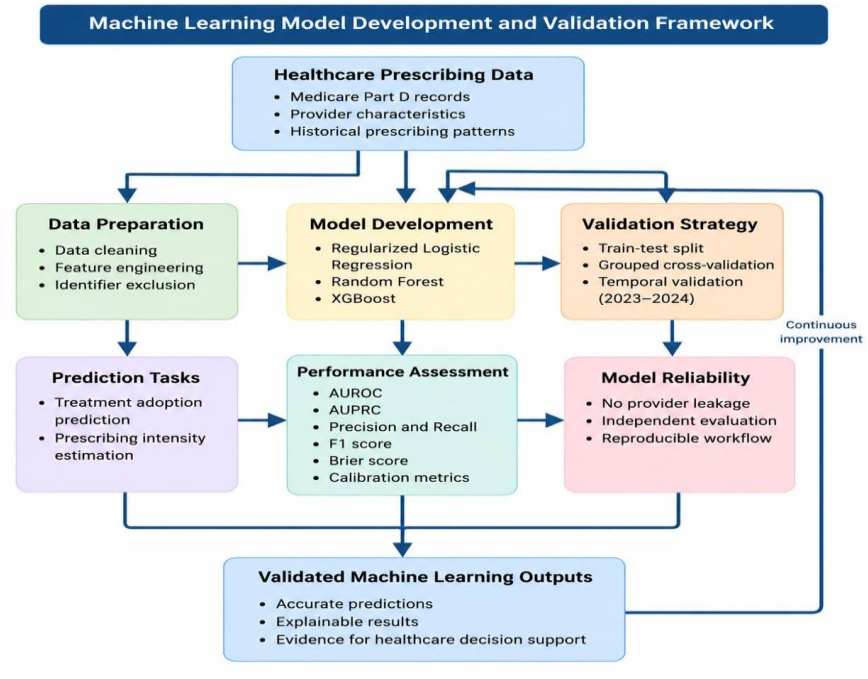


Figure 3: Machine learning development, validation, and evaluation workflow for prescribing behavior prediction.

Figure 3 presents a governed machine-learning framework for predicting treatment adoption by healthcare providers that integrates data governance, predictive modeling, validation, explainability, and monitoring. The framework shows the pipeline from CMS data management and feature control to model prediction and fairness evaluation, and then to SHAP interpretation and human oversight. It highlights how governance mechanisms enable transparent, reliable, and accountable AI deployments in healthcare.

3.4.Explainability, Fairness, and Reproducibility

The study evaluated model explainability using SHAP values from the best-performing tree-based model in the provisional analysis. Other than total claims and standardized fills, candidate predictors included use of older GLP-1s in the prior period, specialty, RUCA category, brand-to-generic prescribing mix, average beneficiary risk score, and the distribution of MAPD versus PDP claims. Researchers retained beneficiaries' race and sex composition for fairness auditing only, not for prediction. Researchers examined model performance separately by specialty, Census region, and urban-rural context, and by beneficiary demographic composition (Patel & Basu, 2026).

The provisional analysis calculated primary discrimination and error measures and estimated 95% bootstrap confidence intervals. To support repeatability in the final execution, the framework specified a versioned data dictionary, fixed random seeds, saved preprocessing parameters, and scripted model execution. These controls were designed so the pipeline could be executed as a repeatable, controlled workflow (Malik, 2024; Koneru, 2025). These procedures addressed perceived gaps and included aspects of the national prescribing study design, including public longitudinal data, temporal validation, explainability, subgroup assessment, and operational governance.

4. Governed Machine Learning Framework

4.1.Data Governance and Provenance Layer

The governance layer maps each output model to a CMS step or source. The framework also recorded metadata for each file added to the pipeline before analysis, including the dataset name, release version, extraction time, source-file checksum, and data dictionary used. The study used NPI only to link provider-level and provider-drug records across years. It also developed a feature lineage register to store the source field, analytical role, model variable, and transformation. Provider drug observations were suppressed separately from true zeros, and each recoding, logarithmic transformation, provider-determined missing-value rule, and derived feature was documented in a transformation log.

The framework specified version-controlled scripts and fixed preprocessing rules to support replication. This aligned with the need for auditable healthcare data pipelines and interoperability in analytics (Rangu et al., 2025). Researchers classified the features into three governance types. Predictors included prescribing rates, specialty class, RUCA classification, prior GLP-1 utilization, beneficiary risk score, prior claims, and MAPD or PDP distribution. Audit variables included only beneficiary race and sex composition, which was used to monitor subgroup performance rather than for prediction (Ravindranath et al., 2025). Excluded fields included NPI, exact address, telephone number, and other fields that could directly identify providers or enable inappropriate profiling.

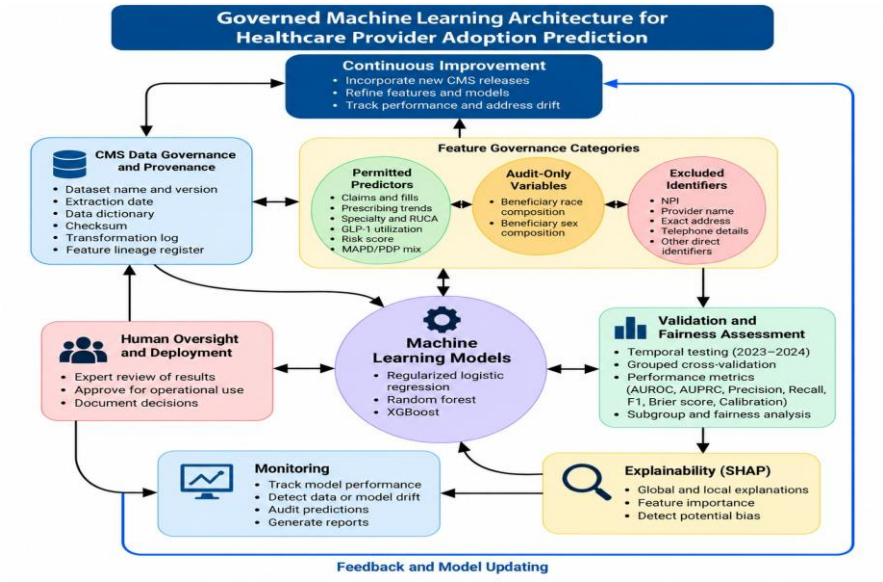


Figure 4: Governed machine learning architecture integrating data governance, prediction, validation, explainability, monitoring, and oversight.

Figure 4 outlines the workflow for creating the machine learning models, building the cohort, generating the features, training the model, testing the model in the temporal cohort, and evaluating model performance. It emphasizes **separating training and test sets**, using grouped cross-validation, and comparing regularized logistic regression, random forest, and XGBoost models. It also summarizes the classification and post-adoption prescribing metrics used to evaluate predictive performance and reliability.

4.2.Prediction and Explainability Layer

The prediction layer transformed the validated CMS records into model-ready features, which were then fed into the logistic regression, random forest, and XGBoost models used in the analysis. Classification produced the probability that a provider would adopt observable semaglutide or tirzepatide prescribing in the next period (Ben Moshe Gaash et al., 2026). Expected standardized 30-day prescribing intensity after adoption was estimated using regression.

The framework requires model probabilities to be calibrated before release, while SHAP values identify the key features driving each prediction. Each governed output thus included an adoption probability, expected prescribing intensity, leading contributing factors, and calibration information. The workflow sequence was CMS data to validation to feature engineering to classifier/regressor to calibration to SHAP explanation to governed output (Rossi et al., 2026). It was not designed to recommend treatment for individual patients or to automatically rank providers for intervention without organizational review.

4.3.Validation, Fairness, and Model Risk Controls

Risk controls were specified before deployment. An AUROC decrease of more than 0.05 from the approved holdout benchmark would trigger an investigation and could lead to model suspension. An expected calibration error greater than 0.05 would trigger recalibration or redevelopment. Researchers checked material drift for any feature with a population stability index greater than 0.20. Subgroup evaluation included specialty, Census region, rurality, and beneficiary composition. The challenger baseline model was kept as a regularized logistic regression model to show that the added model complexity has sufficient merit (Sun, 2026).

The model documentation recorded data restrictions, intended uses, known suppression effects, and limitations. The testing approach could be aligned with the results of workflow-regulated research, including hybrid rule-based validation, which can increase transparency of structured constraints and compliance exceptions (Sepuri, 2026). Together, those controls implement the four principles of fairness, traceability, robustness, and explainability outlined in the FUTURE-AI framework. Table 3 provides information on validation, fairness, and the model risk control plan during and before deployment. The table summarizes the monitoring thresholds, governance actions, and documentation requirements needed to support reliable and responsible healthcare AI use.

Table 3: An overview of validation, fairness, and model risk control measures.

Control Area	Assessment Criteria / Threshold	Governance Action
Predictive performance monitoring	AUROC reduction >0.05 from the approved holdout benchmark	Investigate model degradation and consider suspension or redevelopment

Control Area	Assessment Criteria / Threshold	Governance Action
Calibration monitoring	Expected calibration error >0.05	Perform recalibration or retraining before continued use
Feature stability monitoring	Population Stability Index (PSI) >0.20	Review feature drift and assess impact on model reliability
Fairness evaluation	Performance assessment across specialty, Census region, rurality, and beneficiary demographic composition	Require subgroup analysis before model approval and deployment
Model complexity control	Regularized logistic regression retained as challenger baseline	Confirm that complex models provide measurable improvement over interpretable alternatives
Documentation and transparency	Record intended uses, prohibited uses, dataset limitations, and CMS suppression effects	Maintain complete model documentation and audit trail
Governance principles	Fairness, traceability, robustness, explainability, and monitoring	Ensure alignment with responsible healthcare AI deployment practices

4.4.Deployment, Monitoring, and Human Oversight

The proposed deployment framework was limited to population-level operational support. Potential applications include clinical device monitoring and drug-demand forecasting, planning for provider education, allocating healthcare resources, identifying geographic or specialty areas that could struggle with guideline implementation, and market access planning. Predictions would require interpretation by the appropriate clinical, analytics, or policy personnel before action (Wells et al., 2025). Quarterly monitoring would evaluate feature drift, output drift, discrimination, calibration, subgroup performance, changes in CMS data structure, changes in approved clinical use and reimbursement, and formulary changes. Models would be versioned, with monitoring results stored to maintain an auditable lifecycle that records the model version and monitoring-review outcome (Li et al., 2026).

Human oversight would serve as the final governance mechanism to ensure accountability before any deployment decision. The framework prohibits automated insurance restrictions, automated exclusion of patients from treatment, treatment recommendations for individual patients, and replacement of prescriber judgment. Formal review would be required before continued model use if drift occurred, subgroup performance deteriorated, or treatment indications changed. This governance approach, coupled with provenance, fairness, explainability, monitoring, and accountability for adoption prediction, would turn the statistical adoption-prediction model into a decision-support process closely tied to prediction quality.

5. Results and Discussion

5.1.Sample Characteristics and Treatment Adoption Patterns

The study selected a provisional analytical sample of 120,000 Medicare Part D prescribers. The sample included family practice (34.6%), internal medicine (30.8%), cardiology (22.2%), and endocrinology (12.4%) prescribers. By Census region, 38.3% were in the South, 22.5% in the Midwest, 20.4% in the Northeast, and 18.8% in the West; 14.7% practiced in rural areas. The median Part D prescribing volume across the dataset was 1,482 claims per quarter, with an interquartile range of 678.0 to 3,112.0 claims per quarter. The

median beneficiary volume was 408, and the median HCC risk score and interquartile range (IQR) were 1.19 and 0.94 to 1.98, respectively.

The provisional analysis showed that 41.6% of providers had previously treated patients with GLP-1 therapies. For the independent holdout cohort, 16,487 of 24,000 providers were eligible for observable-adoption assessment at the start of 2024. In the provisional 2024 analysis, 3,129 of the 16,487 eligible providers crossed the CMS reporting threshold, yielding an observable adoption rate of 19.0%. These provisional results show differences by provider characteristics, specialty, healthcare-system context, and prior treatment experience.

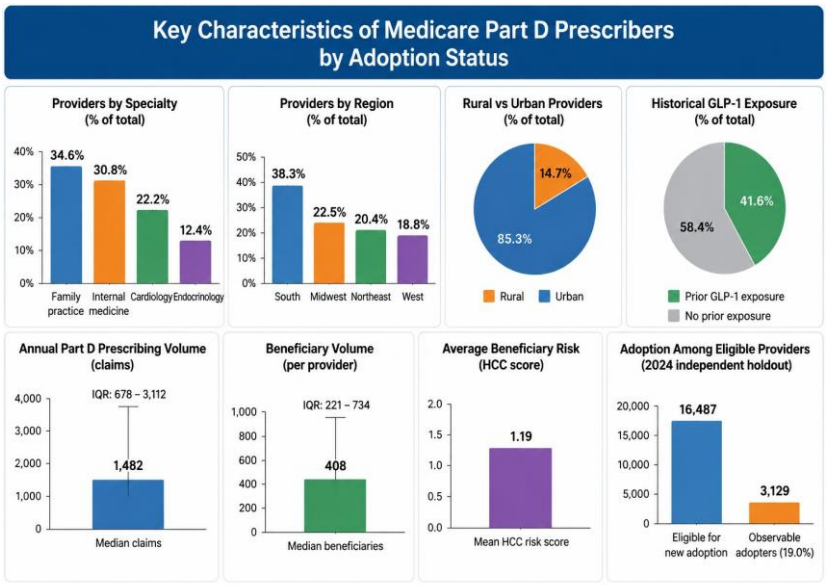


Figure 5: Distribution of key provider characteristics and observable treatment-adoption patterns in the provisional Medicare Part D cohort.

Figure 5 summarizes the provisional cohort by specialty, Census region, rural or urban practice location, previous GLP-1 exposure, prescribing volume, beneficiary volume, beneficiary risk score, and 2024 observable-adoption status. Of the 16,487 adoption-eligible providers, 3,129 crossed the CMS reporting threshold.

5.2.Provisional Treatment Adoption Prediction Performance

In the provisional classification analysis, the more complex models generally performed better. The regularized logistic regression model, an interpretable baseline, had an AUROC of 0.762 and an AUPRC of 0.435. Random forest performed better, with AUROC = 0.817 and AUPRC = 0.536. XGBoost achieved the highest performance, with an AUROC of 0.858 and an AUPRC of 0.617 (Kim et al., 2024). Additional provisional evaluation yielded a Brier score of 0.112, a calibration slope of 0.98, and an approximate F1 value of 0.60. The adoption class had fewer observations than the non-adoption class, making AUPRC particularly informative.

AUPRC better highlights how well the model captures adopters, whereas AUROC evaluates the model's ranking capability across thresholds. Among the provisional results, XGBoost showed the highest performance for discrimination, positive-class recognition, and probability calibration. Moreover, the computing infrastructure and associated governance controls should be scalable to support reliable implementation of such prediction systems (Nagaraj et al., 2025). Table 4 compares the performance of machine learning algorithms

developed to predict treatment adoption in healthcare provider settings. It concisely presents the measures used to evaluate a model's effectiveness: discrimination, positive-class identification, calibration, and predictive reliability.

Table 4: Provisional comparative performance metrics of machine-learning models for treatment-adoption prediction.

Model	AUROC	AUPRC	Additional Evaluation Metrics
Regularized Logistic Regression	0.762	0.435	Interpretable baseline model
Random Forest	0.817	0.536	Improved nonlinear pattern detection
XGBoost	**0.858**	**0.617**	Brier score = 0.112; Calibration slope = 0.98; F1 score ≈ 0.60

Note. AUROC assesses overall ranking performance; when treatment adoption events are infrequent relative to non-adoption, AUPRC better evaluates positive-class identification. In this provisional analysis, XGBoost showed the highest performance for discrimination, adopter identification, and probability calibration.

5.3.Provisional Prescribing Behavior Prediction and Explainability

In the provisional post-adoption prescribing-intensity analysis, XGBoost achieved a mean absolute error of 38.6 standardized 30-day fills and an R^2 of 0.61. The provisional results suggest that providers' future prescribing was strongly associated with their historical practices and that other factors also contributed to prescribing intensity (Ribers & Ullrich, 2024). Key drivers identified by SHAP analysis included prior GLP-1 use, baseline Part D prescribing volume, recent prescribing growth, provider specialty, brand-name prescribing share, beneficiary risk profile, and rural or urban practice context.

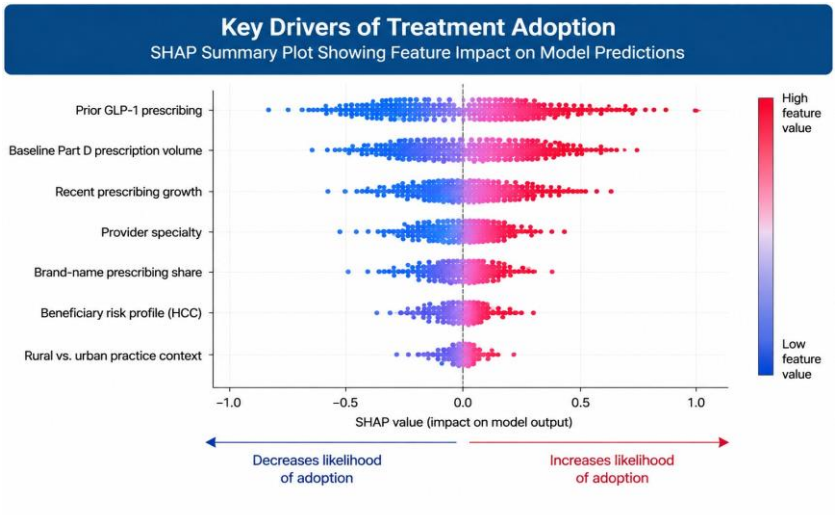


Figure 6: SHAP summary plot showing key predictors influencing provisional treatment-adoption model predictions.

Figure 6 presents a SHAP summary visualization of the relative influence of the leading predictors in the provisional treatment-adoption model. The predictors shown include prior GLP-1 use, baseline Part D prescribing volume, recent prescribing growth, provider specialty, brand-name prescribing share, beneficiary risk profile, and rural or urban practice context.

Predictor effects reflect associations with the model outcome and do not necessarily imply causal influence because other access factors may influence adoption.

5.4.Provisional Fairness, Calibration, and Governance Results

The provisional subgroup evaluation showed generally consistent performance across provider groups. Specialty-specific AUROCs ranged from 0.83 to 0.87, the difference between urban and rural provider AUROCs was less than 0.02, and subgroup ECE values were below 0.05. In addition, the governance-constrained model showed a modest reduction in discrimination when restricted to permitted predictors, excluding direct identifiers and sensitive demographic predictors.

These provisional findings suggest that the governance restrictions may reduce the risk of unnecessary profiling. Model approval analyses also considered performance across specialty, geography, rurality, and beneficiary-group composition (Davoudi et al., 2024). Practices necessary for **reliable artificial intelligence systems** in a governance setting include audit logging, controlled deployment, and periodic performance evaluation (Prakashkumar, 2025).

5.5.Discussion in Relation to the Literature and Real-World Implications

The provisional results are relevant to the current literature because they focus on provider-level treatment uptake using longitudinal administrative prescribing patterns. The provisional XGBoost AUROC of 0.858 was lower than the approximately 0.94 reported in patient-level EHR prediction, potentially because public claims contain fewer clinical variables and the present outcome focuses on provider adoption rather than patient disease prediction. The findings on historical prescribing patterns also align with earlier studies of variation in clinician prescribing by provider characteristics (Abdullah et al., 2024).

The provisional 19.0% observable-adoption rate occurred during a period of substantial growth in semaglutide use and expanding access to incretin therapy. In practice, clinical interest and actual prescribing may diverge because of contractual terms, reimbursement rates, formulary status, and other administrative constraints, all of which may influence observed provider behavior. These findings highlight the need for transparency, monitoring, and governance before deploying AI in healthcare (Saenz et al., 2024). Table 5 reviews results from existing studies and outlines implications for applying machine learning tools in healthcare to predict treatment adoption.

Table 5: Comparison of Provisional Predictive Findings, Literature Evidence, and Real-World Implications.

Discussion Area	Study Findings	Literature/Context Comparison	Real-World Implication
Provider-level treatment adoption prediction	In the provisional analysis, XGBoost achieved an AUROC of 0.858, demonstrating the potential to predict observable provider adoption using longitudinal Medicare Part D prescribing data	Lower than the approximately **0.94 AUROC** reported in patient-level EHR prediction because administrative claims contain fewer clinical variables and focus on provider behavior rather than patient outcomes	Supports the use of provider-level prediction for treatment diffusion monitoring and healthcare planning

Discussion Area	Study Findings	Literature/Context Comparison	Real-World Implication
Role of historical prescribing behavior	Previous prescribing patterns contributed significantly to predicting future adoption	Consistent with evidence that clinician-specific characteristics influence prescribing decisions	Provider experience and prescribing history can improve forecasting of future therapy uptake
Incretin therapy adoption trends	The provisional observable-adoption rate was 19.0% among eligible providers during the 2024 holdout period	Occurred during rapid semaglutide utilization growth and expanding access to incretin therapies	Adoption patterns should be interpreted alongside changing healthcare access conditions
Influence of external access factors	Adoption was not explained solely by provider preference	Reimbursement policies, formulary restrictions, and administrative requirements influence measurable prescribing activity	Models should incorporate access-related factors where possible to distinguish provider behavior from system constraints
Healthcare AI deployment requirements	Provisional predictive performance was combined with transparency, monitoring, and governance controls	Responsible AI principles emphasize explainability, validation, and lifecycle oversight	Healthcare AI systems require governance mechanisms before operational implementation

5.6.Future Research Recommendations

Further research is needed to determine whether the model can be expanded to Medicaid and commercial claims. A multi-institutional electronic health record validation should incorporate additional clinical parameters and constructs and evaluate performance in the clinical environment (Guo et al., 2024). Future models should include plan-specific details on formularies and prior authorization requirements to distinguish access barriers from provider preference. Survival and adoption-likelihood modeling could help describe the diffusion of treatments over time. Peer influence and professional diffusion mechanisms might be identified using provider-network analysis.

Future evaluations should focus on model performance after changes in reimbursement, new clinical indications and supply conditions. Research also needs to determine whether model refinements improve discrimination and support healthcare planning, provider education, and resource allocation (Moreno-Sánchez et al., 2024). Scalable monitoring approaches remain essential for the sustainable deployment of AI systems. Before operational use, prospective external validation should be completed.

6. Conclusions

The researchers developed a governed machine-learning framework to predict treatment uptake and prescribing patterns among healthcare providers using longitudinal Medicare Part D data from 2021 to 2024. It addressed the four objectives by describing provider adoption patterns, comparing predictive models and provider prescribing profiles, and implementing governance controls within the analytical workflow. The study examined provider

characteristics and observable behaviors, including treatment adoption and prescribing intensity, as design factors. Observable adoption was defined as the transition from no published semaglutide or tirzepatide record in the previous year to a published record in the following year. Because CMS suppresses provider-drug counts of 10 or fewer, a published record corresponds to at least 11 claims. This distinction prevented suppressed low-volume prescribing from being interpreted as true nonuse.

The analytical framework showed that provider-level adoption can be analyzed using public data while maintaining model control. The study included 120,000 prescribers: 96,000 for model development and 24,000 for holdout evaluation. In the provisional 2024 holdout analysis, 3,129 of 16,487 eligible providers crossed the observable-adoption threshold, yielding an adoption rate of 19.0%. XGBoost achieved the highest provisional classification performance, with an AUROC of 0.858, an AUPRC of 0.617, a Brier score of 0.112, a calibration slope of 0.98, and an approximate F1 score of 0.60. In the provisional regression analysis, XGBoost achieved a mean absolute error of 38.6 standardized 30-day fills and an R^2 of 0.61. These results remain provisional and should not be interpreted as final empirical results until the complete reproducible analysis has been executed and verified.

The study also demonstrated predictive information in historical prescribing practices. Existing GLP-1 use, Part D volume, recent volume changes, specialty, share of branded prescription volume, beneficiaries' risk profile, and rural or urban practice context were factors that the SHAP analysis identified as predictive of adoption or prescribing intensity. However, these associations were not interpreted as causal explanations of provider behavior. Reimbursement restrictions, formulary access, patient affordability, new indications, and local practice conditions may limit adoption and prescription volumes.

Linking governance with prediction was a central component of the framework. The researchers incorporated data provenance, feature lineage, suppression handling, temporal validation and calibration, subgroup assessment, explainability, challenger models, drift thresholds, and human review. The researchers removed direct identifiers and used beneficiaries' sex and race composition only for fairness auditing. The model was therefore treated as a prediction system for provider behaviors, not a clinical decision-making system. It was not designed to promote or recommend semaglutide, tirzepatide, or any other drug to individual patients, automatically restrict coverage, or discontinue treatment without prescriber input.

The framework provides a solid basis for population-level market access analysis, healthcare resource planning, evidence implementation monitoring, provider education, and pharmaceutical demand forecasting. However, it is based on public CMS claims and suppression rules and has low sensitivity for low-volume prescribers. Thus, external validation is needed and can be achieved through Medicaid, commercial claims, and multi-institutional EHR datasets. Future testing should also account for changes in formularies, reimbursement policies, clinical guidelines, and approved indications. The framework should be externally validated in the future to ensure it is accurate, calibrated, fair, and operationally useful in real-world settings.

References:

Abdullah, M. A., Ahmed, A. S., Shaikh, B. T., Kaleem, M. B., Sumbal, A., & Naeem, Z. (2024). Prescription practices of physicians, quality of care, and patients' safety: A

mixed methods comprehensive study. *International Journal of Health Sciences*, *18*(1), 29–34. https://pmc.ncbi.nlm.nih.gov/articles/PMC10768469/

Ben Moshe Gaash, M., Chodick, G., Romano, R., Greenfeld, S., Gedassi, B., Lerner, U., Azuri, J., & Shapiro Ben David, S. (2026). Physician adoption patterns of AI-driven clinical decision support systems in urinary tract infection management. *Scientific Reports*. https://doi.org/10.1038/s41598-026-57724-7

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

Davoudi, A., Chae, S., Evans, L., Sridharan, S., Song, J., Bowles, K. H., McDonald, M. V., & Topaz, M. (2024). Fairness gaps in machine learning models for hospitalization and emergency department visit risk prediction in home healthcare patients with heart failure. *International Journal of Medical Informatics*, *191*, 105534. https://doi.org/10.1016/j.ijmedinf.2024.105534

Efthimiou, O., Seo, M., Chalkou, K., Debray, T., Egger, M., & Salanti, G. (2024). Developing clinical prediction models: A step-by-step guide. *BMJ*, *386*, e078276. https://doi.org/10.1136/bmj-2023-078276

Enugala, V. K., & Goel, G. (2025). AI-driven sustainable infrastructure and EV supply chain optimization using BIM, LiDAR, and SAP-based automation. In *Proceedings of the 2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)* (pp. 1–7). IEEE. https://ieeexplore.ieee.org/document/11374619

Fourkiotis, K. P., & Tsadiras, A. (2025). Future internet applications in healthcare: Big data-driven fraud detection with machine learning. *Future Internet*, *17*(10), 460. https://doi.org/10.3390/fi17100460

Ghazalbash, S., Zargoush, M., Guilcher, S. J., & Kuluski, K. (2026). Responsible AI for predicting delayed hospital discharge among older adults: Development and evaluation study for balancing accuracy, equity, and explainability. *JMIR Medical Informatics*, *14*, e83244. https://doi.org/10.2196/83244

Gidali, D., & Ward, B. (2023). *The 2021 Physician Pain Management Questionnaire pilot study*. National Center for Health Statistics. https://stacks.cdc.gov/view/cdc/131353

Guo, L. L., Fries, J., Steinberg, E., Fleming, S. L., Morse, K., Aftandilian, C., Posada, J., Shah, N., Sung, L., et al. (2024). A multi-center study on the adaptability of a shared foundation model for electronic health records. *npj Digital Medicine*, *7*, 171. https://doi.org/10.1038/s41746-024-01166-w

Johnphill, O., Sadiq, A. S., Kaiwartya, O., & Taheir, M. A. (2026). Cross: A cloud-native approach to automated remediation and self-healing in cyber-physical systems. *Cybersecurity*, *9*(1), 124. https://doi.org/10.1186/s42400-026-00549-8

Kim, H., et al. (2024). Development of a machine learning approach for prediction of red blood cell transfusion in patients undergoing Cesarean section at a single institution. *Scientific Reports*, *14*, 16556. https://doi.org/10.1038/s41598-024-67784-2

Kondrashov, A. A., Loskutova, E. E., & Kurashov, M. M. (2026). The digital profile of a physician as a new approach to improving recommendation systems for prescribing drugs. *International Journal of Risk & Safety in Medicine, 37*(2), 185–192. https://doi.org/10.1177/09246479251389215

Koneru, N. M. K. (2025). AI-driven optimization of CI/CD pipelines for multi-cloud infrastructure deployment. In *Proceedings of the 2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)* (pp. 1–6). IEEE. https://ieeexplore.ieee.org/abstract/document/11374441

Kyle, M. A., Kakani, P., Burn, S. L., & Maini, L. (2026). Trends in prescription drug coverage restrictions in Medicare, Medicaid, and commercial insurance plans, 2011–2019. *Health Affairs Scholar, 4*(3), qxag059. https://doi.org/10.1093/haschl/qxag059

Lekadir, K., et al. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ, 388*, e081554. https://doi.org/10.1136/bmj-2024-081554

Li, Y., Yuan, M., Yang, Y., Fu, J., Xin, Y., Duan, C. J., & Wang, J. (2026). An ethics-informed computable audit framework for monitoring misdiagnosis risk in AI-assisted diagnosis. *Scientific Reports*. https://doi.org/10.1038/s41598-026-46652-1

Liu, X., Lu, C. A., Shih, Y. T., & Jiang, C. (2025). Coverage and prior authorization policies for semaglutide and tirzepatide in Medicare Part D plans. *JAMA Network Open, 8*(8), e2529842. https://doi.org/10.1001/jamanetworkopen.2025.29842

Lulla, K., & Sayyed, Z. (2025). AI-driven fault-tolerant cloud-native architecture for mission-critical hardware-software integration and scalable enterprise automation. In *Proceedings of the 2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)* (pp. 1–6). IEEE. https://ieeexplore.ieee.org/document/11374324

Malik, G. (2024). Embedding security into the pipeline: A framework for scalable DevSecOps implementation. *International Journal of Data Science and Machine Learning, 4*(1), 37–62. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=QeYN5fsAAAAJ&citation_for_view=QeYN5fsAAAAJ:UeHWp8X0CEIC

Malik, G., & Prashasti. (2023). Continuous exposure management using AI and threat intelligence. *International Journal of Intelligent Systems and Applications in Engineering, 11*(6s). https://scholar.google.com/citations?view_op=view_citation&hl=en&user=QeYN5fsAAAAJ&citation_for_view=QeYN5fsAAAAJ:Se3iqnhoufwC

Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Navarro, C. A., Dhiman, P., et al. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ, 388*, e082505. https://doi.org/10.1136/bmj-2024-082505

Moreno-Sánchez, P. A., Aalto, M., & van Gils, M. (2024). Prediction of patient flow in the emergency department using explainable artificial intelligence. *Digital Health, 10*. https://doi.org/10.1177/20552076241264194

Nagaraj, V., Gundla, S. R., & Vennamaneni, P. R. (2025). AI-driven secure and cloud-native microservice architectures for high-performance financial and hardware accelerator systems. In *Proceedings of the 2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)* (pp. 1–7). IEEE. https://ieeexplore.ieee.org/document/11374740

Patel, S. Y., & Basu, S. (2026). Integrating healthcare system context to improve risk prediction and assess racial disparities among dual-eligible Medicare–Medicaid beneficiaries: A retrospective cohort study using national fee-for-service claims. *BMJ Open, 16*(3), e111343. https://bmjopen.bmj.com/content/16/3/e111343

Prakashkumar, P. K. R. (2025). Analytical study of adapting Oracle AI Agent Studio into Oracle ERP overview. *International Journal of Entrepreneurship, Innovation, and Business Strategies, 3*(1), 90–103. https://gprjournals.org/journals/index.php/ijeib/article/view/452

Rangu, S., Chadha, K. S., & Samala, S. (2025). AI-powered cloud architecture, agile analytics, and enterprise DevOps-enabled framework for healthcare and insurance ecosystem optimisation. In *Proceedings of the 2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)* (pp. 1–7). IEEE. https://ieeexplore.ieee.org/document/11374414

Ravindranath, R., Stein, J. D., Hernandez-Boussard, T., Fisher, A. C., Wang, S. Y., Amin, S., et al. (2025). The impact of race, ethnicity, and sex on fairness in artificial intelligence for glaucoma prediction models. *Ophthalmology Science, 5*(1), 100596. https://doi.org/10.1016/j.xops.2024.100596

Ribers, M. A., & Ullrich, H. (2024). Complementarities between algorithmic and human decision-making: The case of antibiotic prescribing. *Quantitative Marketing and Economics, 22*, 445–483. https://doi.org/10.1007/s11129-024-09284-1

Rossi, D., Auriemma Citarella, A., De Marco, F., Di Biasi, L., Zheng, H., & Tortora, G. (2026). DREAM: Diabetes risk via explainable AI modeling. *Multimedia Tools and Applications, 85*(2), 145. https://doi.org/10.1007/s11042-026-21240-7

Saenz, A. D., Centi, A., Ting, D., You, J. G., Landman, A., Mishuris, R. G., et al. (2024). Establishing responsible use of AI guidelines: A comprehensive case study for healthcare institutions. *npj Digital Medicine, 7*, 348. https://doi.org/10.1038/s41746-024-01300-8

Sardana, J., Hariharan, R., & Pudi, R. S. (2025). AI-driven secure and resilient healthcare data platforms with intelligent search and personalized communication. In *Proceedings of the 2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)* (pp. 1–6). IEEE. https://ieeexplore.ieee.org/document/11374370

Sargent, M., Matthews, L. J., Yon, G. V., Storholm, E. D., Ober, A. J., & Green, H. D., Jr. (2024). Assessing the dynamics of PrEP adoption in a national-scale physician network. *Social Networks, 78*, 226–237. https://doi.org/10.1016/j.socnet.2024.02.001

Scannell, C., Romley, J., Myerson, R., Goldman, D., & Qato, D. M. (2024). Prescription fills for semaglutide products by payment method. *JAMA Health Forum, 5*(8), e242026. https://doi.org/10.1001/jamahealthforum.2024.2026

Sepuri, S. (2026). Comparative analysis of artificial intelligence test generation methods for compliance validation in regulated enterprise workflows. *Journal of Intelligent Decision Making and Information Science, 3*(2s), 107–123. https://doi.org/10.59543/jidmis.v3.632

Sun, K. (2026). Machine learning-based prediction model for 30-day emergency department revisits in a medically underserved tertiary hospital: Formative retrospective cohort study. *JMIR Formative Research, 10*, e87289. https://doi.org/10.2196/87289

Wells, B. J., Nguyen, H. M., McWilliams, A., Pallini, M., Bovi, A., Kuzma, A., et al. (2025). A practical framework for appropriate implementation and review of artificial intelligence (FAIR-AI) in healthcare. *npj Digital Medicine, 8*(1), 514. https://doi.org/10.1038/s41746-025-01900-y

Wesolowicz, D. M., Becker, W. C., Alegria, M., Chin-Purcell, L., Glass, J. E., Knudsen, H. K., et al. (2025). Integrating implementation outcomes into effectiveness studies: A practical guide for clinical interventionists. *Journal of General Internal Medicine, 40*(10), 2376–2382. https://doi.org/10.1007/s11606-025-09551-1

Yaseen, A., & Lahiri, S. W. (2023). Health care provider prescribing habits and barriers to use of new type 2 diabetes medications: A single-system survey study. *Clinical Diabetes, 41*(4), 490–501. https://doi.org/10.2337/cd22-0100

Zhang, Y. W., Blyumin, M. G., Qu, C., Guo, X., & Zhang, M. (2026). Assessing the impact of the Inflation Reduction Act on Medicare prescription drug coverage. *American Journal of Health-System Pharmacy*, zxag083. https://doi.org/10.1093/ajhp/zxag083