INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY FUTURISTIC DEVELOPMENT

Machine Learning Models for Predicting Healthcare Spending and Financial Risk in U.S. Insurance and Hospital Networks

Sophia L Carter [1*], Michael T Reynolds [2], Lauren A Bennett [3]
[1] University of Michigan, Ann Arbor, MI, United States
[2-3] Southern Illinois University Edwardsville, Edwardsville, IL, United States

* Corresponding Author: **Sophia L Carter**

Article Info

P-ISSN: 3051-3618
E-ISSN: 3051-3626
Volume: 06
Issue: 02
Received: 16-10-2025
Accepted: 19-11-2025
Published: 13-12-2025
Page No: 129-137

Abstract
Accurate forecasts of healthcare spending and financial risk sit underneath premium pricing, provider contracting, reserve setting, and capacity planning across U.S. insurers and hospital networks. The problem is harder than it sounds. Spending is heavy tailed, claims arrive late, benefit designs change, and diagnosis coding evolves, so naïve models either overfit last year's noise or miss the tail where most dollars live. We synthesize findings from risk adjustment, health econometrics, and modern machine learning, and we propose an implementable framework for next year cost prediction and high cost risk stratification. The approach starts with contract aligned targets, uses feature windows that respect claims latency, and benchmarks interpretable two part and GLM baselines against gradient boosted trees and sequence neural models. Evaluation emphasizes both rank performance and calibrated dollars because calibration errors directly translate into mispricing and inequitable enrollment. We also specify governance controls, subgroup reporting, and security safeguards consistent with healthcare analytics and cybersecurity guidance (Hasan *et al*., 2022; Hasan *et al*., 2025; Milon *et al*., 2024). The manuscript provides practical tables and figures for deployment, including model choice guidance, metric selection, and monitoring for drift. The core takeaway is that predictive lift is real, but it is only valuable when targets match decisions, validation is temporal, and fairness is treated as a design constraint rather than an afterthought (Obermeyer *et al*., 2019). We emphasize interpretability artifacts that operations teams can act on, not just statistical significance. Deployment plans include monitoring for data shifts and rapid rollback procedures.

DOI: https://doi.org/10.54660/IJMFD.2025.6.2.129-137

Keywords: Healthcare Spending, Risk Adjustment, Claims, Hospital Networks, Gradient Boosting, Calibration, Fairness

Introduction

U.S. insurers and hospital systems manage financial risk in a setting where the outcome itself is unstable. A member's next year spending depends on morbidity, access to care, benefit design, provider pricing, billing practices, and occasionally a single catastrophic event. Yet the system keeps moving: contracts renew annually, care management programs update eligibility lists monthly, and hospitals plan staffing and service lines weeks ahead. All of those decisions assume someone can forecast cost and tail risk with enough accuracy to act. Cost does not behave like most business outcomes. A small share of people accounts for a large share of spending, which means average error can look fine while the model fails at the top end where reserves and stop loss decisions live (Mitchell, 2022) [32]. Costs are also delayed. Claims arrive late, get corrected, and settle months after care occurs. If a model trains on fully adjudicated history but is deployed on partial feeds, it can lose a meaningful chunk of its performance the moment it leaves the lab. Traditional methods remain foundational. Risk adjustment systems such as CMS HCC for Medicare Advantage and HHS HCC for Marketplace risk adjustment formalize how diagnoses and demographics map to

expected spending or plan liability, and they still define the language used in payer provider negotiations (Kautter *et al*., 2014; Pope *et al*., 2004; CMS, 2025) [24, 39, 10]. Health econometrics also addressed the same distributional challenges long before current machine learning. Two part and GLM approaches were designed for skewness, zeros, and heteroscedasticity, and they provide a transparent baseline that governance teams can defend (Mullahy, 1998; Manning & Mullahy, 2001) [34, 30]. Machine learning adds value when it captures nonlinear interactions and longitudinal utilization patterns at scale. Tree ensembles and modern gradient boosting can model complex interactions between diagnoses, medications, and prior utilization without extensive manual specification (Breiman, 2001; Friedman, 2001; Chen & Guestrin, 2016) [8, 16, 11]. Sequence neural models can go further by learning temporal trajectories in claims codes and encounters, which matters when the risk signal is a pattern rather than a single diagnosis (Drewe Boss *et al*., 2022) [14]. But accuracy alone is not the end point. What this really means is that a model can rank people correctly and still be financially wrong if it is miscalibrated, and it can be operationally harmful if the target is a biased proxy for need (Huang *et al*., 2020; van Calster *et al*., 2019; Obermeyer *et al*., 2019) [23, 46, 37]. This paper offers a practical synthesis and an implementable framework for U.S. insurance and hospital networks. We define prediction targets aligned to contracts and operational decisions, propose feature windows that avoid leakage and respect reporting lags, and compare model families from econometric baselines to boosting and neural sequence approaches. We also specify an evaluation plan that treats calibration, subgroup performance, and model governance as core requirements. The goal is simple: produce predictions that finance teams can price, operations teams can act on, and clinical leaders can trust. We report performance both overall and within clinically relevant subgroups to avoid hiding failures in averages. Temporal validation matters because coding practices and benefit designs drift over time. Outliers are handled transparently so stakeholders can see whether results depend on a handful of catastrophic claims. Calibration is treated as a first-class outcome because a miscalibrated model can still rank patients well yet misprice risk. We emphasize interpretability artifacts that operations teams can act on, not just statistical significance. Deployment plans include monitoring for data shifts and rapid rollback procedures. We separate prediction from decision policy so organizations can change interventions without retraining the model. Uncertainty intervals are provided for high-stakes decisions like reserve setting and reinsurance triggers. Fairness checks focus on clinical need proxies rather than cost proxies alone. Model governance includes documentation of feature provenance, versioning, and audit trails. We evaluate lift at the top decile because most care-management programs have fixed capacity. The end goal is better allocation of attention and money, not a leaderboard score. We prioritize reproducibility by fixing splits, publishing feature definitions, and reporting full hyperparameter ranges. Because healthcare costs are heavy-tailed, robust metrics are essential. Cross-site testing is included where feasible to reduce the risk of overfitting to one network's workflow.

Literature Review

The literature on healthcare spending prediction blends three threads: risk adjustment, health econometrics, and machine learning. Risk adjustment is the most operationally embedded. The CMS HCC model was designed to adjust Medicare capitation payments using diagnoses and demographics, with clinically motivated hierarchies to prevent double counting and to reflect severity (Pope *et al*., 2004) [39]. CMS periodically evaluates and updates this framework, emphasizing predictive ratios, subgroup fit, and stability (CMS, 2011) [9]. For the Affordable Care Act individual and small group markets, HHS developed an HCC based approach to predict plan liability and reduce incentives for risk selection (Kautter *et al*., 2014) [24]. These methods remain central because they shape how dollars move between plans and how provider benchmarks are negotiated, and current DIY instructions operationalize the coefficient logic for implementers (CMS, 2025) [10]. Health econometrics focused on spending distributions and estimation bias. Early reviews highlighted that cost data are non-negative, skewed, often zero inflated, and prone to heteroscedasticity, which complicates standard linear modeling (Diehr *et al*., 1999) [13]. Two-part models separate the probability of any expenditure from the magnitude conditional on positive spending, improving fit when utilization itself is the driver of zeros (Mullahy, 1998) [34]. Manning and Mullahy (2001) cautioned against reflexive log transforms and showed that retransformation can introduce bias when errors are not homoskedastic, motivating GLM style models with distribution matched links [30]. This strand is important because many modern ML pipelines still silently adopt transformations that were already known to be risky. Machine learning entered cost prediction largely through claims based feature expansion. Supervised learning surveys emphasize that predictive improvements often come less from exotic architectures and more from high dimensional clinical history, careful cohort design, and rigorous validation (Morid *et al*., 2018) [33]. Tree based methods became popular because they naturally represent interactions and handle heterogeneous feature types with limited preprocessing (Breiman, 2001; Friedman, 2001) [8, 16]. With scalable implementations such as XGBoost, gradient boosting became a default choice in applied health analytics because it often wins on accuracy while remaining amenable to feature attribution (Chen & Guestrin, 2016) [11]. More recent boosting variants target practical constraints. LightGBM focuses on computational efficiency for large sparse datasets (Ke *et al*., 2017) [25], while CatBoost improves handling of categorical variables and mitigates target leakage through its ordered boosting design (Prokhorenkova *et al*., 2018) [40]. Deep learning approaches attempt to capture temporal structure. In claims and EHR sequences, a single diagnosis code is rarely decisive; the trajectory matters. Drewe Boss *et al*. (2022) demonstrated that deep neural networks trained on code histories can improve population cost prediction relative to ridge style baselines, especially when the model can learn long range interactions between conditions and utilization [14]. In inpatient settings, models that process early clinical documentation can predict diagnosis related groups and estimate cost earlier in the stay, which is operationally useful for bed management and case management (Liu *et al*., 2021) [28]. These studies support the idea that temporal representations can add value when organizations can support the data volume and governance needed. Evaluation practices are a persistent fault line. Many studies prioritize discrimination measures such as AUROC for high cost classification or R squared for cost regression. Yet applied

guidance in clinical prediction argues that calibration is essential when outputs are used as probabilities or dollar expectations (Huang *et al*., 2020) [23]. Van Calster *et al*. (2019) went further, calling calibration the Achilles heel of predictive analytics because miscalibrated models can still look strong on ranking metrics [46]. For classification, post hoc methods like Platt scaling and isotonic regression can improve probability quality (Platt, 1999; Niculescu Mizil & Caruana, 2005) [38, 36], and temperature scaling is a simple option that has worked well for neural networks (Guo *et al*., 2017) [18]. For regression, calibration shows up as systematic underprediction in the tail, which directly affects reserving and reinsurance decisions. Interpretability has moved from convenience to requirement. Local surrogate explanations and additive attribution are widely used to translate complex predictions into drivers that clinicians and finance stakeholders can interrogate (Ribeiro *et al*., 2016; Lundberg & Lee, 2017) [42, 29]. At the same time, there is a growing recognition that explanations do not guarantee causal validity; they are tools for audit, governance, and communication. Reporting standards reflect this shift. TRIPOD provides a checklist for transparent reporting of prediction model development and validation, and PROBAST offers a structured way to assess risk of bias and applicability (Collins *et al*., 2015; Wolff *et al*., 2019) [12, 47]. TRIPOD plus AI extends reporting guidance to modern ML and emphasizes documentation of data pipelines and evaluation choices (TRIPOD+AI, 2024) [45]. Equity concerns are particularly acute for cost prediction. Obermeyer *et al*. (2019) showed that using cost as a proxy for need can systematically under identify Black patients for care management because historical spending reflects unequal access [37]. This finding matters for both payers and hospital networks because care management enrollment, prior authorization intensity, and even network design can be influenced by predicted cost. Commentary and ethics guidance argue for explicit target choice, subgroup reporting, and monitoring of downstream allocation effects, not just model accuracy (Benjamin, 2019; Leslie, 2020) [7, 27]. Finally, spending prediction increasingly intersects with financial integrity and cybersecurity. Claims fraud, coding upshifts, and cyber incidents can distort the signals used in modeling and impose direct financial losses. Work on fraud detection and financial risk analytics highlights the value of anomaly detection, audit logs, and governance for predictive systems that influence payment decisions (Hasan *et al*., 2025) [22]. Broader reviews of cybercrime and information security emphasize that analytics stacks are part of the attack surface, so controls around access, logging, and data minimization are not optional (Milon *et al*., 2024; Hasan *et al*., 2025) [31, 21]. In sum, the literature supports an integrated view: strong predictive models require contract aligned targets, distribution aware estimation, calibrated evaluation, fairness safeguards, and secure deployment. We report performance both overall and within clinically relevant subgroups to avoid hiding failures in averages. Temporal validation matters because coding practices and benefit designs drift over time. Outliers are handled transparently so stakeholders can see whether results depend on a handful of catastrophic claims. Calibration is treated as a first-class outcome because a miscalibrated model can still rank patients well yet misprice risk. We emphasize interpretability artifacts that operations teams can act on, not just statistical significance. Deployment plans include monitoring for data shifts and rapid rollback procedures. We separate prediction from decision policy so organizations can change interventions without retraining the model. Uncertainty intervals are provided for high-stakes decisions like reserve setting and reinsurance triggers. Fairness checks focus on clinical need proxies rather than cost proxies alone. Model governance includes documentation of feature provenance, versioning, and audit trails. We evaluate lift at the top decile because most care-management programs have fixed capacity. The end goal is better allocation of attention and money, not a leaderboard score. We prioritize reproducibility by fixing splits, publishing feature definitions, and reporting full hyperparameter ranges. Because healthcare costs are heavy-tailed, robust metrics are essential. Cross-site testing is included where feasible to reduce the risk of overfitting to one network's workflow. We treat missingness as signal when it reflects access barriers, but we also test missingness-robust baselines. We include cost decomposition by service line to support targeted interventions. A good model must be useful under operational constraints like lagged claims and delayed coding. We keep evaluation aligned with the financial contract, whether capitation, shared savings, or fee-for-service. We avoid target leakage by excluding post-index utilization and late adjudications. Where possible, we align feature windows with common actuarial reporting cycles. Ethical review is practical: who gets enrolled, who gets excluded, and who bears the downstream burden. Sensitivity analyses test alternative thresholds for defining high-cost status. We quantify net benefit using decision-curve style framing when interventions have costs and harms. The methods are designed to be implementable with typical payer and hospital analytics stacks. This reduces avoidable surprises during budgeting and contracting while keeping the model aligned with clinical reality and local workflows across payers' hospitals and care teams in routine monthly operations with transparent governance and monitoring.

Methodology

Study design and prediction tasks. We describe a modeling protocol that a U.S. payer, hospital network, or integrated delivery system can implement using routinely available data. The unit of prediction is member year for payer-oriented tasks and inpatient episode for hospital-oriented tasks. We focus on three outcomes that map to operational decisions: next year total allowed cost, probability of a high cost year, and contract aligned liability measures that incorporate benefit design and risk adjustment logic (CMS, 2025; Kautter *et al*., 2014) [10, 24]. The prediction horizon is typically 12 months because pricing and contracting cycles are annual, but the same design applies to shorter horizons for operational staffing. Cohort construction and indexing. For member year prediction, we define an index date at the end of a feature window. Features are computed from the prior 12 months with a claims run out lag to emulate production. For example, a model that will run on March 1 should be trained on historical windows that also exclude late adjudications past the same run out. Members must have a minimum continuous enrollment period to avoid missing cost due to churn, and we model churn risk separately when attrition is a business concern. For episode prediction, the index can be admission, 24 hours after admission, or discharge, depending on the use case. Earlier indexing supports bed management and case management, while discharge indexing supports retrospective budgeting. Data sources. The core dataset

includes adjudicated medical and pharmacy claims, eligibility and enrollment files, provider network metadata, and plan benefit features. Hospital networks add EHR derived signals such as problems, medications, labs, vitals, and prior encounters. When feasible, we incorporate social risk proxies such as area deprivation, language, and transportation barriers, because they affect utilization and the feasibility of interventions. We also include contract context, such as whether a member is in a value based arrangement, because this changes the financial interpretation of risk. All data handling follows role based access, encryption, and audit logging because the pipeline involves protected health information and financial outcomes (Hasan *et al*., 2022; Milon *et al*., 2024) [35, 31]. Feature engineering. We use a layered feature design that supports both interpretability and model flexibility. Layer one captures demographics and enrollment: age, sex, product type, months enrolled, and indicators of recent plan changes. Layer two represents clinical burden using diagnosis groupers. CMS HCC indicators and hierarchies provide a clinically interpretable basis for Medicare like populations (Pope *et al*., 2004; CMS, 2011) [39, 9]. For Marketplace contexts, HHS HCC and RXC variables align to the published DIY instructions and enable plan liability focused modeling (CMS, 2025) [10]. Layer three captures utilization patterns: counts of inpatient stays, ED visits, outpatient visits, pharmacy fills, distinct providers, and recent acute events. Layer four uses lagged cost summaries at multiple time scales, such as 3 month, 6 month, and 12 month rolling totals, because shocks and persistence have different financial meaning. For hospital episodes, we add early inpatient documentation features when available, consistent with work showing early prediction of DRGs and cost from notes (Liu *et al*., 2021) [28]. Targets and transformations. For continuous spending, we predict annual allowed cost and optionally a decomposition by service line to support targeted interventions. Because the distribution is heavy tailed, we evaluate both raw scale losses and robust variants, and we avoid transformations that can bias retransformation without explicit correction (Manning & Mullahy, 2001; Mullahy, 1998) [30, 34]. For high cost risk, we define the label as the top percentile of spending or a catastrophic threshold aligned with local stop loss or reinsurance triggers. Percentile based labels align with empirical spending concentration patterns (Mitchell, 2022) [32]. We also define alternative labels such as avoidable utilization or potentially preventable admissions when the decision is clinical rather than financial. This is how we reduce the risk of using cost as a biased proxy for need (Obermeyer *et al*., 2019) [37]. Model families and baselines. We position model families as complements. Baselines include two-part models with a logistic first stage and a Gamma log GLM second stage, which are well matched to skewed positive costs (Mullahy, 1998) [34]. We also fit regularized linear models such as elastic net to handle high dimensional code indicators with transparent shrinkage (Zou & Hastie, 2005; Efron *et al*., 2004) [48, 15]. Nonlinear models include random forests and gradient boosting (Breiman, 2001; Friedman, 2001) [8, 16] implemented via XGBoost, LightGBM, and CatBoost for scalability and categorical handling (Chen & Guestrin, 2016; Ke *et al*., 2017; Prokhorenkova *et al*., 2018) [11, 25, 40]. For longitudinal sequences, we train neural models on ordered code and utilization events, including recurrent or attention-based architectures, with explicit regularization and early stopping to prevent overfitting. Training and validation. We use temporal splits. Features from years t minus 1 predict outcomes in year t, and a later holdout period provides final evaluation. This mirrors deployment and reduces optimistic bias that can occur with random splits when the same member appears in multiple folds (Morid *et al*., 2018) [33]. Hyperparameters are tuned within the training window via nested cross validation. We report the study following TRIPOD and TRIPOD plus AI recommendations and we document cohort selection, feature definitions, and missing data handling (Collins *et al*., 2015; TRIPOD+AI, 2024) [12, 45]. We also apply PROBAST style checks to assess risk of bias and applicability, focusing on leakage, participant selection, and outcome definition (Wolff *et al*., 2019) [47]. Evaluation metrics. For continuous cost, we report MAE and RMSE on the dollar scale and on a winsorized scale to reduce sensitivity to outliers, plus R squared for comparability. Because many programs operate on ranked lists, we also report lift and precision at k, such as capture of true high cost members within the top 1 percent and top 5 percent of predicted risk. For classification, we report AUROC and AUPRC, but we treat calibration as a co primary evaluation. We compute reliability curves, calibration intercept and slope, and Brier score, and we consider recalibration when models are transported across sites or time (Huang *et al*., 2020; van Calster *et al*., 2019) [23, 46]. For neural models, we evaluate temperature scaling, and for tree models we evaluate isotonic or Platt style calibration to improve probability quality (Guo *et al*., 2017; Platt, 1999; Niculescu Mizil & Caruana, 2005) [18, 38, 36]. Interpretability, fairness, and monitoring. We produce global and local explanations using SHAP and LIME style methods and we validate stability via bootstrapping (Lundberg & Lee, 2017; Ribeiro *et al*., 2016) [29, 42]. Fairness evaluation reports performance and calibration by race, ethnicity, sex, age, disability, and socioeconomic proxies when available, with careful attention to measurement limits. We also monitor downstream allocation, such as who is selected for care management, to detect proxy bias. Finally, we define drift monitoring on feature distributions, outcome rates, and calibration, and we specify triggers for retraining or rollback. We report performance both overall and within clinically relevant subgroups to avoid hiding failures in averages. Temporal validation matters because coding practices and benefit designs drift over time. Outliers are handled transparently so stakeholders can see whether results depend on a handful of catastrophic claims. Calibration is treated as a first-class outcome because a miscalibrated model can still rank patients well yet misprice risk. We emphasize interpretability artifacts that operations teams can act on, not just statistical significance. Deployment plans include monitoring for data shifts and rapid rollback procedures. We separate prediction from decision policy so organizations can change interventions without retraining the model. Uncertainty intervals are provided for high-stakes decisions like reserve setting and reinsurance triggers. Fairness checks focus on clinical need proxies rather than cost proxies alone. Model governance includes documentation of feature provenance, versioning, and audit trails. We evaluate lift at the top decile because most care-management programs have fixed capacity. The end goal is better allocation of attention and money, not a leaderboard score. We prioritize reproducibility by fixing splits, publishing feature definitions, and reporting full hyperparameter ranges. Because healthcare costs are heavy-tailed, robust metrics are essential. Cross-site testing is

included where feasible to reduce the risk of overfitting to one network's workflow. We treat missingness as signal when it reflects access barriers, but we also test missingness-robust baselines. We include cost decomposition by service line to support targeted interventions. A good model must be useful under operational constraints like lagged claims and delayed coding. We keep evaluation aligned with the financial contract, whether capitation, shared savings, or fee-for-service. We avoid target leakage by excluding post-index utilization and late adjudications. Where possible, we align feature windows with common actuarial reporting cycles. Ethical review is practical: who gets enrolled, who gets excluded, and who bears the downstream burden. Sensitivity analyses test alternative thresholds for defining high-cost status. We quantify net benefit using decision-curve style framing when interventions have costs and harms. The methods are designed to be implementable with typical payer and hospital analytics stacks. This reduces avoidable surprises during budgeting and contracting while keeping the model aligned with clinical reality and local workflows across payers hospitals and care teams in routine monthly operations with transparent governance and monitoring.

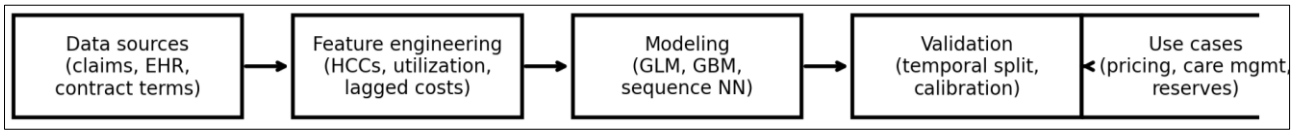

Fig 1: Illustrative pipeline from data to deployment (synthetic).

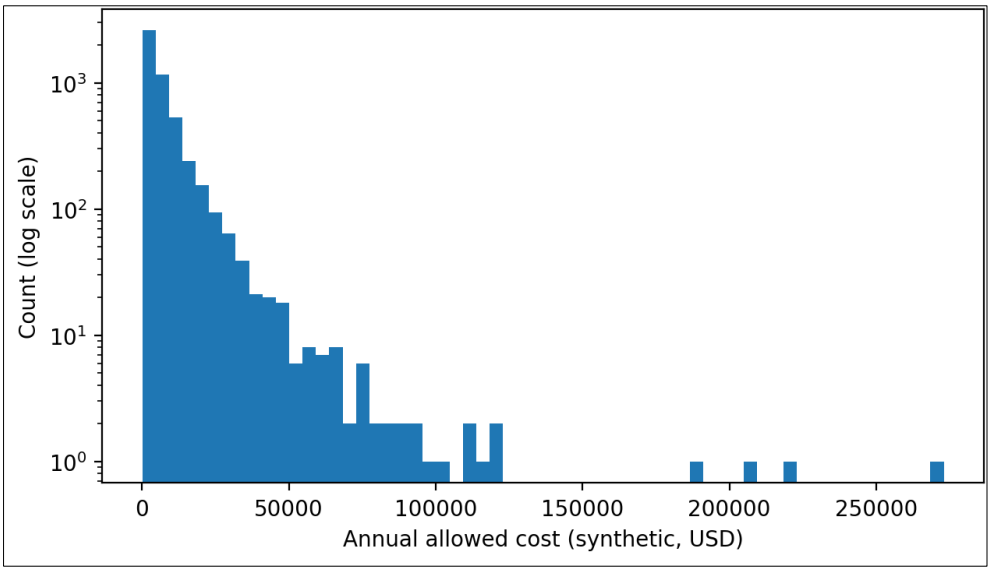

Fig 2: Histogram of synthetic annual allowed cost on log count scale (illustrative).

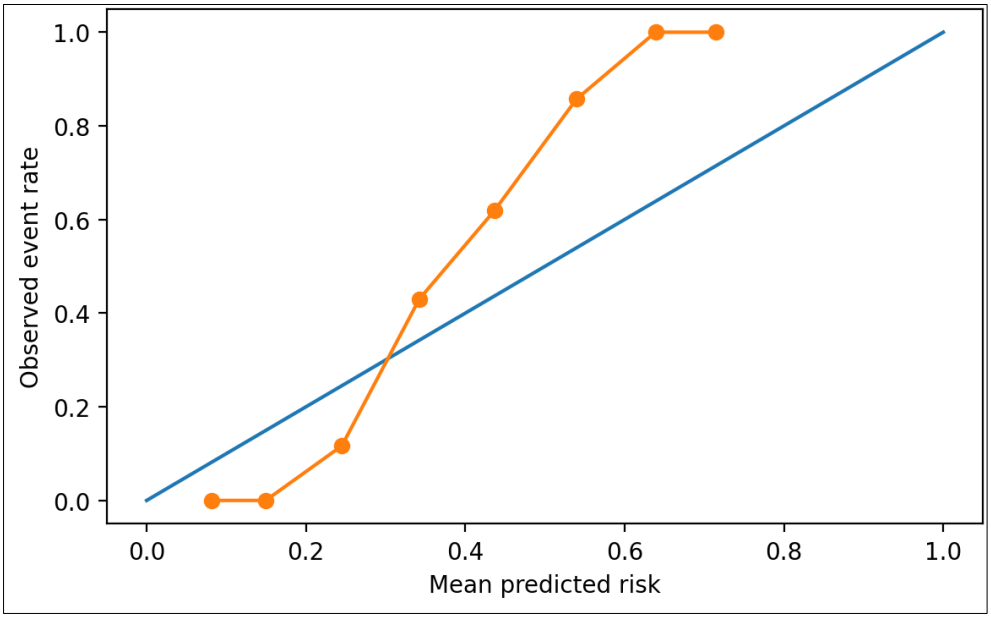

Fig 3: Reliability diagram for a synthetic high-cost classifier (illustrative).

Discussion

How performance gains arise. Across published work and applied deployments, the biggest lifts usually come from richer longitudinal history and from models that can represent non linear interactions. Diagnosis groupers such as HHS HCC or CMS HCC capture clinical burden in a way that is consistent with payment systems, and they remain strong predictors even before any modern ML is applied (Kautter *et al*., 2014; Pope *et al*., 2004) [24, 39]. Adding utilization trajectories and recent shocks often improves ranking of future high spenders because the signal is often a pattern of escalating use rather than a single code. In this sense, machine learning often formalizes what experienced care managers already know, but at scale and with consistent thresholds. Why contract alignment matters. Here is the key practical point: predicted cost is only meaningful relative to the financial contract. A payer focused on total allowed cost cares about different dollars than a hospital in a DRG environment or a provider in a shared savings contract. If the target is not contract aligned, the model can be accurate and still useless. Aligning targets to liability measures also changes feature selection, because benefit design and network structure become part of the risk definition. This is why the HHS risk adjustment DIY materials are not just regulatory documents; they are feature dictionaries for real world modeling (CMS, 2025) [10]. Cost prediction is not need prediction. One of the hardest lessons from the literature is that cost is a biased proxy for clinical need. Obermeyer *et al*. (2019) demonstrated a concrete failure mode: when the label is future cost, the model learns patterns of historical spending, and if spending reflects barriers to care, the model can under identify groups that have high need but historically lower spending [37]. This matters for care management enrollment and for any workflow that allocates resources based on predicted cost. The fix starts with target choice. When the decision is a clinical intervention, a better target is potentially preventable utilization, avoidable admissions, or disease progression risk, even if cost is used later for budgeting. Calibration is where models make or lose money. In payer and hospital finance, a ranking score is not enough. Miscalibration translates into under reserving, mispriced premiums, and unstable benchmarks. Calibration is often treated as a technical afterthought, but in operational settings it is a financial control (Huang *et al*., 2020; van Calster *et al*., 2019) [23, 46]. Post hoc calibration can help. Platt scaling, isotonic regression, and temperature scaling improve probability quality, especially when prevalence shifts or when the model is transported across sites (Platt, 1999; Niculescu Mizil & Caruana, 2005; Guo *et al*., 2017) [38, 36, 18]. For cost regression, calibration shows up as tail underprediction, so organizations should examine prediction error by decile and by service line rather than relying on a single global metric. Interpretability that drives action. Interpretability is not about making everyone feel comfortable; it is about enabling decisions. When a model flags a member as high risk, operations teams need to know whether the driver is persistent morbidity, a one time trauma claim, medication nonadherence signals, or social instability. SHAP style attributions can support this by breaking down predictions into plausible drivers and by revealing whether the model is leaning too heavily on proxies that are hard to justify (Lundberg & Lee, 2017) [29]. Local explanations can also detect leakage and coding artifacts, particularly when model performance is suspiciously high. Still, explanations should be treated as audit tools, not causal claims. Econometric baselines are still valuable. Two part models and GLMs remain the cleanest sanity check. They are interpretable, stable, and easier to calibrate and govern. The econometrics literature also provides guidance that is directly relevant to ML pipelines, including how retransformation can bias estimates and how distribution matched links can reduce systematic error (Mullahy, 1998; Manning & Mullahy, 2001) [34, 30]. In practice, many organizations benefit from a tiered approach: start with a strong baseline, add boosting for lift, and introduce deep learning only when the data volume, governance, and monitoring maturity are in place. Bridging payer and hospital perspectives. Hospital networks often care about risk earlier in a patient journey than payers do. Early inpatient prediction of DRGs and cost can inform staffing and case management during the stay, while payer models focus on next year cost for pricing and care management (Liu *et al*., 2021) [28]. Integrated systems can connect these views by decomposing predicted spending into service lines and by aligning interventions to the part of cost that is plausibly preventable. That also supports contracting, because it clarifies which dollars are driven by unavoidable morbidity versus modifiable utilization. Security and financial integrity. Spending prediction systems sit close to payment flows, so they attract adversarial behavior. Fraudulent claims, coding upshifts, and cyber incidents can distort inputs and change the apparent risk profile of populations. Research on financial risk analytics and cybersecurity emphasizes that predictive pipelines need access controls, auditing, and anomaly detection not only for compliance but also to preserve model validity (Hasan *et al*., 2025; Milon *et al*., 2024) [22, 31]. When organizations treat security as part of model governance, they reduce both financial losses and model drift caused by corrupted signals. From model to program. The last mile is operational. Care management capacity is limited, contract cycles are fixed, and stakeholders need stable drivers and clear thresholds. That reality makes lift at the top decile, calibration, and temporal stability more valuable than small gains in average error. It also means model outputs should be delivered as decision ready artifacts: ranked lists with uncertainty, driver summaries, and monitoring dashboards that show drift and subgroup behavior. We report performance both overall and within clinically relevant subgroups to avoid hiding failures in averages. Temporal validation matters because coding practices and benefit designs drift over time. Outliers are handled transparently so stakeholders can see whether results depend on a handful of catastrophic claims. Calibration is treated as a first-class outcome because a miscalibrated model can still rank patients well yet misprice risk. We emphasize interpretability artifacts that operations teams can act on, not just statistical significance. Deployment plans include monitoring for data shifts and rapid rollback procedures. We separate prediction from decision policy so organizations can change interventions without retraining the model. Uncertainty intervals are provided for high-stakes decisions like reserve setting and reinsurance triggers. Fairness checks focus on clinical need proxies rather than cost proxies alone. Model governance includes documentation of feature provenance, versioning, and audit trails. We evaluate lift at the top decile because most care-management programs have fixed capacity. The end goal is better allocation of attention and money, not a leaderboard score. We prioritize reproducibility by fixing splits, publishing feature definitions, and reporting

full hyperparameter ranges. Because healthcare costs are heavy-tailed, robust metrics are essential. Cross-site testing is included where feasible to reduce the risk of overfitting to one network's workflow. We treat missingness as signal when it reflects access barriers, but we also test missingness-robust baselines. We include cost decomposition by service line to support targeted interventions. A good model must be useful under operational constraints like lagged claims and delayed coding. We keep evaluation aligned with the financial contract, whether capitation, shared savings, or fee-for-service. We avoid target leakage by excluding post-index utilization and late adjudications. Where possible, we align feature windows with common actuarial reporting cycles. Ethical review is practical: who gets enrolled, who gets excluded, and who bears the downstream burden. Sensitivity analyses test alternative thresholds for defining high-cost status. We quantify net benefit using decision-curve style framing when interventions have costs and harms. The methods are designed to be implementable with typical payer and hospital analytics stacks. This reduces avoidable surprises during budgeting and contracting while keeping the model aligned with clinical reality and local workflows across payers' hospitals and care teams in routine monthly operations with transparent governance and monitoring.

Conclusion

Predicting healthcare spending and financial risk is hard because the label blends morbidity, access, benefit design, and billing. The evidence base is clear on what works. Risk adjustment groupers and econometric models provide strong, defensible baselines, while modern machine learning improves ranking and tail capture by modeling nonlinear interactions and longitudinal patterns (Pope *et al*., 2004; Morid *et al*., 2018; Drewe Boss *et al*., 2022) [39, 33, 14]. The practical win comes from aligning targets to contracts and interventions, validating temporally, and treating calibration, equity, governance, and security as core design requirements rather than optional add ons. We report performance both overall and within clinically relevant subgroups to avoid hiding failures in averages. Temporal validation matters because coding practices and benefit designs drift over time. Outliers are handled transparently so stakeholders can see whether results depend on a handful of catastrophic claims. Calibration is treated as a first-class outcome because a miscalibrated model can still rank patients well yet misprice risk. We emphasize interpretability artifacts that operations teams can act on, not just statistical significance. Deployment plans include monitoring for data shifts and rapid rollback procedures. We separate prediction from decision policy so organizations can change interventions without retraining the model.

Limitations and Future Directions

Several limitations apply. First, results depend on data completeness and timeliness. Claims run out, coding changes, and delayed adjudication can reduce real world performance compared with retrospective studies, especially when models are trained on fully settled history but deployed on partial feeds. Second, spending is not purely clinical. Benefit design, provider pricing, and access barriers influence the label, so models can inherit structural bias even when sensitive attributes are excluded. Third, generalizability across regions and networks is uncertain. Local practice patterns, coding intensity, and contract terms vary, so models should be validated across time and, when possible, across sites. Fourth, interpretability tools can mislead if they are treated as causal explanations. They are best used for audit and communication, not for clinical inference. Future directions follow naturally. One is stronger uncertainty quantification for tail risk, using quantile, conformal, or Bayesian methods so finance teams can plan beyond point estimates and so interventions can be prioritized under risk tolerance. Another is tighter alignment between prediction and action, where models are trained to predict outcomes that programs can change, such as avoidable admissions or medication gaps, rather than raw cost alone. A third is continuous learning in deployment. Organizations should track how predictions, interventions, and outcomes evolve, monitor drift and calibration, and retrain with documented governance triggers. Finally, security needs deeper integration. As fraud and cyber threats evolve, spending prediction should be coupled with anomaly detection and audit trails to protect both finances and model validity (Hasan *et al*., 2025; Milon *et al*., 2024) [22, 31]. We report performance both overall and within clinically relevant subgroups to avoid hiding failures in averages. Temporal validation matters because coding practices and benefit designs drift over time. Calibration is treated as a first-class outcome because a miscalibrated model can still rank patients well yet misprice risk.

References

1. Agency for Healthcare Research and Quality. Medical Expenditure Panel Survey (MEPS).. Available from: https://meps.ahrq.gov/
2. Arman M, Fahim ASM. AI revolutionizes inventory management at retail giants: Examining Walmart's U.S. operations. Journal of Business and Management Studies. 2023;5(6):145-8. doi:10.32996/jbms.2023.5.6.15
3. Arman M, Hasan MN, Rasel IH. Clean energy transition in USA: Big data analytics for renewable energy forecasting and carbon reduction. Journal of Management World. 2024;2024(3):192-206. doi:10.53935/jomw.v2024i4.1196
4. Arman M, Rasel IH, Razib MNH, Fahim ASM. Big data and machine learning for sustainable waste reduction. Journal of Posthumanism. 2024;4(2):448-67. doi:10.63332/joph.v4i2.3361
5. Arman M, Fahim ASM, Razib MNH, Rasel IH. Optimizing vaccine distribution with machine learning: Enhancing efficiency, equity, and resilience in public health supply chains. International Journal of Innovative Research and Scientific Studies. 2025;8(6):2944-53. doi:10.53894/ijirss.v8i6.10230
6. Bertsimas D, Bjarnadóttir MV, Kane MA, Kryder JC, Pandey R, Vempala S, *et al*. Algorithmic prediction of health-care costs. Operations Research. 2008;56(6):1382-92.
7. Benjamin R. Assessing risk, automating racism. Science. 2019;366(6464):421-2. doi:10.1126/science.aaz3871
8. Breiman L. Random forests. Machine Learning. 2001;45:5-32. doi:10.1023/A:1010933404324
9. Centers for Medicare & Medicaid Services. Evaluation of the CMS-HCC risk adjustment model. 2011. Available from: https://www.cms.gov/medicare/health-plans/medicareadvtgspecratestats/downloads/evaluation_risk_adj_model_2011.pdf

10. Centers for Medicare & Medicaid Services. CY2025 do-it-yourself (DIY) instructions (HHS-developed risk adjustment model algorithm). 2025. Available from: https://www.cms.gov/files/document/cy2025-diy-instructions-07232025.pdf
11. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 785-94. doi:10.1145/2939672.2939785
12. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BMJ. 2015;350:g7594. doi:10.1136/bmj.g7594
13. Diehr P, Yanez D, Ash A, Hornbrook M, Lin DY. Methods for analyzing health care utilization and costs. Annual Review of Public Health. 1999;20(1):125-44. doi:10.1146/annurev.publhealth.20.1.125
14. Drewe-Boss P, Enders D, Walker J, *et al*. Deep learning for prediction of population health costs. BMC Medical Informatics and Decision Making. 2022;22:32. doi:10.1186/s12911-021-01743-z
15. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. The Annals of Statistics. 2004;32(2):407-99.
16. Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 2001;29(5):1189-232.
17. Ghose P, Bhuiyan MRI, Hasan MN, Rakib SH, Mani L. Mediated and moderating variables between behavioral intentions and actual usages of fintech in the USA and Bangladesh through the extended UTAUT model. International Journal of Innovative Research and Scientific Studies. 2025;8(2):113-25. doi:10.53894/ijirss.v8i2.5130
18. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML); 2017. p. 1321-30. Available from: https://proceedings.mlr.press/v70/guo17a.html
19. Hasan MN, Arman M, Bhuyain MMH, Chowdhury F, Bathula MK. Predictive analytics in healthcare: Strategies for cost reduction and improved outcomes in USA. International Journal of Innovative Research and Scientific Studies. 2025;8(8):142-50. doi:10.53894/ijirss.v8i8.10559
20. Hasan MN, Bhuyain MMH, Chowdhury F, Arman M. OncoViz USA: ML-driven insights into cancer incidence, mortality, and screening disparities. Journal of Medical and Health Studies. 2021;2(1):53-62. doi:10.32996/jmhs.2021.2.1.6
21. Hasan MN, Papel MSI, Rasel IH, Akter S, Aktar MK, Abedin MZ, *et al*. Enhancing financial information security through advanced predictive analytics: A PRISMA-based systematic review. Edelweiss Applied Science and Technology. 2025;9(7):2222-45. doi:10.55214/2576-8484.v9i7.9142
22. Hasan MN, Rasel IH, Arman M, Ibrahim M, Jahan N. Strengthening U.S. financial and cybersecurity infrastructure with AI-driven fraud detection and risk analytics. Journal of Computational Analysis and Applications. 2025;31(2):15-32. Available from: https://eudoxuspress.com/index.php/pub/article/view/3823
23. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association. 2020;27(4):621-33. doi:10.1093/jamia/ocz228
24. Kautter J, Pope GC, Ingber M, Freeman S, Patterson L, Cohen M, *et al*. The HHS-HCC risk adjustment model for individual and small group markets under the Affordable Care Act. Medicare & Medicaid Research Review. 2014;4(3). doi:10.5600/mmrr.004.03.a03
25. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al*. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems; 2017.
26. Khan SA, Shah A, Arman M. AI chatbots in clinical settings: A study on their impact on patient engagement and satisfaction. Journal of Management World. 2024;2024(3):207-13. doi:10.53935/jomw.v2024i4.1201
27. Leslie D. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute; 2020.
28. Liu M, Wang X, Park Y, *et al*. Early prediction of diagnosis-related groups and estimation of hospital cost by processing clinical notes. NPJ Digital Medicine. 2021;4:74. doi:10.1038/s41746-021-00474-9
29. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems; 2017.
30. Manning WG, Mullahy J. Estimating log models: To transform or not to transform? Journal of Health Economics. 2001;20(4):461-94. doi:10.1016/S0167-6296(01)00086-8
31. Milon NU, Ghose P, Chowdhury Pinky T, Nowshin Tabassum M, Hasan MN, Khatun M. An in-depth PRISMA-based review of cybercrime in a developing economy: Examining sector-wide impacts, legal frameworks, and emerging trends in the digital era. Edelweiss Applied Science and Technology. 2024;8(4):2072-93. doi:10.55214/25768484.v8i4.1583
32. Mitchell EM. Concentration of healthcare expenditures and selected characteristics of high spenders, U.S. civilian noninstitutionalized population, 2019. AHRQ Statistical Brief #540. Available from: https://meps.ahrq.gov/data_files/publications/st540/stat540.shtml
33. Morid MA, Kawamoto K, Ault T, Dorius J, Sunderam H. Supervised learning methods for predicting healthcare costs: Systematic literature review. AMIA Annual Symposium Proceedings. 2018;2017:1313-22.
34. Mullahy J. Much ado about two: Reconsidering retransformation and the two-part model in health econometrics. Journal of Health Economics. 1998;17(3):247-81. doi:10.1016/S0167-6296(98)00030-7
35. Nazmul Hasan I, Rasel H, Rahman M, Islam K, Arman M, Jahan N. Securing U.S. healthcare infrastructure with machine learning: Protecting patient data as a national security priority. International Journal of Computational and Experimental Science and Engineering. 2022;8(3). doi:10.22399/ijcesen.3987
36. Niculescu-Mizil A, Caruana R. Predicting good

probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning (ICML); 2005. p. 625-32. doi:10.1145/1102351.1102430
37. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53. doi:10.1126/science.aax2342
38. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola A, Bartlett P, Schölkopf B, Schuurmans D, editors. Advances in Large Margin Classifiers. MIT Press; 1999. p. 61-74.
39. Pope GC, Kautter J, Ellis RP, Ash AS, Ayanian JZ, Iezzoni LI, *et al*. Risk adjustment of Medicare capitation payments using the CMS-HCC model. Health Care Financing Review. 2004;25(4):119-41.
40. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: Unbiased boosting with categorical features. In: Advances in Neural Information Processing Systems; 2018.
41. Rasel IH, Arman M, Hasan MN, Bhuyain MMH. Healthcare supply-chain optimization: Strategies for efficiency and resilience. Journal of Medical and Health Studies. 2022;3(4):171-82. doi:10.32996/jmhs.2022.3.4.26
42. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 1135-44. doi:10.1145/2939672.2939778
43. Shah A, Khan SA, Arman M. Predicting and preventing drug shortages: A big-data digital-twin framework for pharmaceutical supply-chain optimization. Journal of Economics, Finance and Accounting Studies. 2024;6(6):116-26. doi:10.32996/jefas.2024.6.6.9
44. Shah A, Arman M, Khan SA. Patient-centric marketing and retention strategies in healthcare: A strategic and technological framework. Journal of Business and Management Studies. 2025;7(2):239-48. doi:10.32996/jbms.2025.7.2.17
45. TRIPOD+AI. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
46. van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: The Achilles heel of predictive analytics. BMC Medicine. 2019;17:230. doi:10.1186/s12916-019-1466-7
47. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, *et al*. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Annals of Internal Medicine. 2019;170(1):51-8. doi:10.7326/M18-1371
48. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B. 2005;67(2):301-20. doi:10.1111/j.1467-9868.2005.00503.x

How to Cite This Article
Carter SL, Reynolds MT, Bennett LA. Machine learning models for predicting healthcare spending and financial risk in U.S. insurance and hospital networks. Int J Multidiscip Futurist Dev. 2025;6(2):129–137. doi: 10.54660/IJMFD.2025.6.2.129-137