International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5776 - 5785
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0750

Predictive Modeling of HEDIS Measure Performance for Intervention Prioritization in Medicare Advantage Star Ratings: A Methodological Framework

Sachin Girdhar[1]
[1]*Kurukshetra University, Kurukshetra (Haryana), India*
Email id: *s.girdhar88@gmail.com*[1]

Abstract

The Centers for Medicare & Medicaid Services (CMS) Five-Star Quality Rating System evaluates Medicare Advantage (MA) plans on an annual basis, and measures based on the Healthcare Effectiveness Data and Information Set (HEDIS) make up a significant portion of the rating system and are used to qualify for quality bonus payments and greater rebate retention. Outreach and clinical capacity are limited and plans must select which open care gaps to address. Conventional gap-closure programs prioritize members based on the presence of a gap, which is not a good indicator of the likelihood of contact affecting the outcome being measured. To aggregate methodological evidence to support prediction of HEDIS measure compliance at the individual level and suggest a feasible system of translating calibrated predictions into an expected-value intervention priority score for MA quality programs.Narrative review and framework synthesis based on published evaluations of MA Star Ratings and pay-for-performance, and program documentation from CMS and the National Committee for Quality Assurance, as well as the prediction-model methodology literature. Results were categorized into the steps a plan analytics team must take, sequentially, from cohort building to deployment. The five criteria that distinguish useful models from operationally inert ones are the following: denominators and labels aligned to certified HEDIS specifications; temporally honest validation mirroring the deployment point; calibration rather than discrimination as the primary criterion; separation of the predicted residual gap from predicted responsiveness to intervention; and subgroup auditing, since label prevalence differs by social risk. A priority score is specified, measure level and model class considerations are tabulated, measure weight, expected uplift, residual gap probability and feasibility/cost. Predictive modeling can only enhance the performance of Star Ratings when the estimand is the same as the decision. Plans should include a model of incremental gap closure under a capacity constraint, validation of calibration within social-risk strata, and testing of deployed models against a concurrent counterfactual.

Keywords: *HEDIS; Medicare Advantage; Star Ratings; predictive modeling; machine learning; quality improvement; care gap closure; health equity*

1. Introduction

Medicare beneficiary is enrolled in Medicare Advantage, and the Five-Star Quality Rating System is the government's primary tool to reward or penalize plan quality. Four-star and higher contracts are eligible for quality bonus payments and receive a higher percentage of rebate dollars to invest in supplemental benefits; the rating is also prominently featured during open enrollment and is linked to enrollment choice [1,2]. Financial impacts of a fractional change in a summary rating are thus high, and they are asymmetric around the administrative thresholds due to the step function nature of the rating with respect to the underlying measure scores [1,3,4]. There are systematic differences in the rated performance of contracts, depending on the characteristics of the provider organizations, including their size, tax status and tenure [5] and the broader pay-for-performance literature suggests that provider organizations respond to such incentives, but the size and persistence of the response is variable [7]. The largest single block of measures in the Part C rating is derived from HEDIS measures maintained by the National Committee for Quality Assurance and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5776-5785
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0750

includes cancer screening, diabetes and cardiovascular care, care coordination and care for older adults [6]. The program has two components that influence the analytical problem. Firstly, measures are weighted differently, with process measures typically receiving a weight of 1, intermediate outcome measures a weight of 3, and outcome measures a weight of 5 [1]. Second, the thresholds for each of the star levels are determined by clustering the distribution of contract performance and thus the value of an incremental numerator hit is determined by a contract's distance from the nearest cut point, not on the absolute score it achieves [1,4]. In this context, most plans have gap-closure programs that produce lists of members whose measure gaps are open and send those members to telephonic outreach or mailings, incentive offers, and/or home-based services or provider-directed reporting. The logic used for the dominant list generation is descriptive: find all the members eligible who have not yet been met by the numerator, possibly sorted by the number of open gaps. This logic combines three different possibilities: Do gaps exist today? Would they exist at the end of the measurement year if there were no contact? Would contact alter that? The third quantity is the only decision value when capacity is binding. This article aims to discuss the methodological needs to estimate these quantities from the routinely available plan data and to define a framework that transforms these quantities into a defensible prioritization rule.

2. Methods

This is a narrative review using framework synthesis and is reported as such. Three bodies of literature were identified: methodological literature on the development, validation and reporting of clinical prediction models; empirical literature on MA Star Ratings, pay-for-performance and quality measurement, including literature on social risk and measure performance; and applied machine learning literature related to tabular administrative claims data. Program specifications were taken from CMS Star Ratings technical documentation and rulemaking, and from NCQA HEDIS measure resources [1,4,6]. Sources were chosen for their methodological relevance to the prioritization decision and the framework has been structured around the series of decisions an analytics team would make when building a cohort, from definition to deployment and evaluation.

3. The decision problem: from gap volume to expected value

Allow a member i to be eligible for measure m in a particular measurement year. There are three relevant quantities. The probability that fails the numerator at year-end under current practice and no additional outreach is called the residual gap probability ($\hat{p}$gap). The incremental probability of numerator compliance due to a specific intervention is the expected uplift ($\hat{u}$). The marginal value of a numerator hit depends upon the measure weight w and the location of the contract relative to the closest cut point in the measure. None of these are the target of the common practice of ranking by gap presence. It also over-invests in two groups: members who would have closed the gap without contact, and members with structural barriers that are unlikely to be overcome by a reminder call. Both mistakes can be prevented with information plans already in hand. Probable self-closers can be distinguished from persistent non-compliers with reasonable accuracy on most process measures with the following information: Pharmacy claims, prior-year measure history, encounter continuity, and interaction history with the plan. The uplift component is more challenging and is where most programs fall short. There is evidence on the effectiveness of reminder and recall interventions, with meaningful, but heterogeneous effects, and responses focussed in subgroups defined by prior engagement and access [8]. Heterogeneity suggests that a single average effect estimate is insufficient for targeting and that plans with outreach at scale should use their own historical campaigns as a source of quasi-experimental variation to estimate conditional effects. Table 1 provides an overview of measure-level properties that affect the actual behavior of these quantities. Each measure has a different qualifying look-back period, completeness and timeliness of the data used to determine numerator compliance, and provider documentation and specialist supply requirements to close out the measure. Such differences are not

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5776-5785
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0750

coincidental. They decide whether or not to predict at all for a measure and what the model should predict. As shown in Table 1 Characteristics of representative HEDIS-derived and related measures in the Medicare Advantage Star Ratings program and their implications for predictive modeling.

Table 1 Characteristics of representative HEDIS-derived and related measures in the Medicare Advantage Star Ratings program and their implications for predictive modeling

Measure (abbreviation)	Measure type	Typical weight	Primary data source	Principal modeling and operational implication
Breast Cancer Screening (BCS-E)	Process	1	Administrative claims; ECDS	Long (roughly two-year) qualifying window means prior-year screening status dominates prediction; capacity and appointment seasonality limit late-year closure.
Colorectal Cancer Screening (COL-E)	Process	1	Claims; supplemental data	Multiple modalities with different look-back periods; for mailed stool-based kits the useful prediction target is kit return, not screening per se.
Eye Exam for Patients With Diabetes (EED)	Process	1	Claims; supplemental vendor files	Specialist supply and travel distance are strong predictors; in-office retinal imaging changes feasibility more than member motivation does.
Glycemic Status Assessment for Patients With Diabetes (GSD)	Intermediate outcome	3	Laboratory feeds; EHR	Depends on clinical result capture; result latency shortens the actionable window and biases mid-year gap lists.
Blood Pressure Control (CBP)	Intermediate outcome	3	EHR or hybrid/ECDS abstraction	Requires clinical data interoperability; measurement variability adds label noise that attenuates apparent model performance.
Statin Therapy for Patients With Cardiovascular Disease (SPC)	Process	1	Pharmacy and medical claims	Pharmacy signals are timely and near-complete, making this among the most predictable measures; therapy refusal and statin intolerance are the residual drivers.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5776-5785
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0750

Plan All-Cause Readmissions (PCR)	Outcome	3	Claims	Risk-adjusted and ratio-based; small contract denominators produce unstable estimates and weak individual-level discrimination.
Transitions of Care (TRC); Follow-Up after Emergency Department Visit (FMC)	Process	1	Claims; admission-discharge-transfer feeds	Numerator windows of 7 to 30 days require near-real-time event notification; prediction is useful mainly for triaging which discharges need staffed follow-up.
Medication adherence, diabetes / hypertension / cholesterol (Part D, PDC ≥ 80%)	Intermediate outcome (Part D)	3	Pharmacy claims	Refill history supports accurate early-year prediction and an explicit mathematical ceiling on whether the threshold remains attainable.
Care for Older Adults (COA)	Process	1	Supplemental data; ECDS	Closure depends on documentation workflow rather than member behavior; provider-level features carry most of the signal.

Abbreviations: ECDS, Electronic Clinical Data Systems; EHR, electronic health record; PDC, proportion of days covered. Weights shown are the categories that have applied in recent rating years; measure inventories, weights and specifications are revised annually and must be verified against the CMS Technical Notes and NCQA HEDIS volume applicable to the rating year under study [1,6].

4. Data foundations and feature construction

Most operational models break down at the cohort construction stage, and it's not a modelling problem, it's a specification problem. The denominator should reflect the certified HEDIS denominator logic for the measure, which includes continuous enrollment requirements, age and sex restrictions, history of an event or diagnosis, and exclusions for advanced illness and frailty or for cancer screening, prior bilateral mastectomy or total colectomy [6]. If the denominator is not exactly what is modeled, then a model trained on the approximate denominator will provide scores that will not match the denominator on which the plan is ultimately measured, and an inability to reconcile scores will erode the credibility of the analytics function with quality operations staff faster than any error in the model. The label must be tied to the same year-end determination as the audited rate and the information included in the feature set must be information that is available at the year-end determination date. Plans use gap lists in the first or second quarter, so a model that is validated with features as seen up to December will be much higher than the actual performance. The proper discipline is to set a scoring date, censor all features on that date and train separately for each deployment window in case lists are updated. Prior year compliance on the same and related measures, utilization intensity, continuity of primary care, pharmacy fill patterns (summarized as proportion of days covered, gap-in-therapy indicators), density

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5776-5785
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0750

of specialist and diagnostic capacity in the member's area, attributed provider and practice identifiers, dual eligibility and low-income subsidy status and prior response to plan contact are all features of domains that consistently carry signal. When individual-level social data are not available, area-level socioeconomic measures can also provide information that is predictive of claims, and they are widely used for this purpose [9,10] but are a coarse proxy for individual circumstances and thus have implications for use, discussed in section 7. There are two important practical considerations to highlight. Much of the clinical information, such as for blood pressure control or care for older adults, comes from supplemental data feeds, and these feeds are incomplete and laggy: models that assume they are informative missingness will learn the data pipeline and not the clinical phenomenon. Sample size for model development should not be assumed to be large due to large membership; sample size is the denominator for the measure of interest and events of interest may be rare in some measures [11].

5. Model how to use classes and their appropriate use

Gradient-boosted decision trees tend to be best at discrimination, and are relatively tolerant of missingness and non-linearity in tabular administrative data, such as that contained in plans. Penalized regression is still useful when the feature space is wide and sparse and/or a transparent, auditable specification is desired. Simple logistic regression still has a place as a reference model: If a complex learner cannot outperform a well-specified regression by an amount that is operationally significant, the extra governance load is not worth it [13,14]. In cases where the question is whether or not something has happened, continuous-time survival formulations are more appropriate, but if the question is when, then discrete-time formulations are more appropriate, and they treat disenrollment as censoring instead of a missing outcome. Multilevel models with provider random effects provide the provider-level quantities that network and value-based contracting teams need and stabilize estimates when there is variation in performance due to practices, not providers. When the question is who to contact, the estimand is a conditional treatment effect and uplift or causal machine learning methods estimate the conditional treatment effect directly. The comparisons of these classes are shown in Table 2. In this context, the requirement of interpretability is not a philosophical one, but rather a practical one. Care managers don't act on a score unless they can view why a member surfaced and compliance functions require that scores not be based on variables a plan cannot defend with. The Shapley-value attribution offers local explanations, is now common practice for tree ensembles [15] and can be used for both purposes. However, explanations must not be made in the form of causal statements about individual members. As shown in Table 2. Comparison of model classes for predicting HEDIS measure compliance and prioritizing intervention in Medicare Advantage populations.

Table 2 Comparison of model classes for predicting HEDIS measure compliance and prioritizing intervention in Medicare Advantage populations

Model class	Typical role in HEDIS gap prediction	Strengths	Limitations and cautions
Logistic regression with splines	Reference model and regulatory-facing baseline	Transparent coefficients; well-behaved calibration; simple to govern and audit	Requires manual specification of interactions; underfits when predictors number in the hundreds

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5776-5785
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0750

Penalized regression (LASSO, elastic net)	High-dimensional claims and pharmacy feature sets	Automatic variable selection; stable in wide, sparse data; retains interpretability	Selection is unstable among correlated features; coefficients should not be read causally
Gradient-boosted trees (XGBoost, LightGBM)	Primary predictive engine for most process measures	Best discrimination in tabular claims data; handles missingness and non-linearity	Probabilities often need recalibration; sensitive to distribution shift; needs post hoc explanation
Random forest	Screening for candidate features and interactions	Low tuning burden; robust to noise	Probability estimates compressed toward the mean; large memory footprint at scale
Multilevel (hierarchical) models	Separating member-level from provider- and site-level variation	Partial pooling stabilizes small panels; supports provider feedback reports	Computationally heavy with many clusters; convergence problems in sparse panels
Discrete-time survival models	Timing of gap closure within the measurement year	Answers when, not only whether; accommodates censoring at disenrollment	More complex label construction; requires monthly person-period data
Uplift and causal machine learning (T-learner, X-learner, causal forest)	Estimating who responds to outreach rather than who has a gap	Directly targets incremental effect; avoids spending capacity on likely self-closers	Needs randomized or quasi-experimental outreach data; estimates are noisier than risk models
Contextual bandits	Sequential choice of outreach channel and timing	Learns online; adapts to changing response rates	Requires disciplined logging and governance; difficult to validate retrospectively

Abbreviation: HEDIS, Healthcare Effectiveness Data and Information Set. Model classes are complementary rather than mutually exclusive; a common production configuration pairs a gradient-boosted risk model with a separately estimated uplift model and a transparent regression benchmark.

6. A prioritization framework

The framework is provided in Figure 1. Panel A illustrates the steps from data assembly to the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5776-5785
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0750

construction of the labels, feature engineering, model fitting and calibration to a composite priority score, deployment under the capacity constraint, and evaluation, which is again used for the next training. The priority score for member i and measure m is the product of the measure weight, the expected uplift, the residual gap probability and a feasibility term, divided by the expected cost of the intervention. The feasibility term reflects whether closure is still possible within the calendar year, which for adherence measures can be calculated exactly from the arithmetic of the proportion of days covered, and which for screening measures is based on the availability of appointments. Panel B depicts the targeting logic for one measure. The two axes are the predicted residual gap and the expected uplift and the product defines iso-priority contours. Intensive outreach is targeted at members with a high residual gap and high expected uplift. Members who have a high residual gap but low expected uplift generally have structural or clinical obstacles and should be referred for case management, transportation or in-home services instead of another reminder call. Low residual gap probability members should only be contacted for an automated reminder at low cost, as most will do what is required without contact. This two dimensional view reveals an explicit tradeoff that is not made in single score gap lists. There are two aggregation problems when the framework is used on the entire measure set. First, the marginal value of a numerator hit depends on the contract and the threshold; a numerator hit on a measure that is just below a cut point has a temporarily higher marginal value, while a numerator hit on a measure that is well above the four-star threshold has little marginal value. Distance weighting is valid but should be capped as cut points vary from year to year and are not known a priori [1,4]. Second, when a single contact can address multiple gaps, the members who are eligible for multiple measures should be prioritized based on the sum of their priority scores at the measure level, as the marginal cost of an additional gap within a single contact is small.

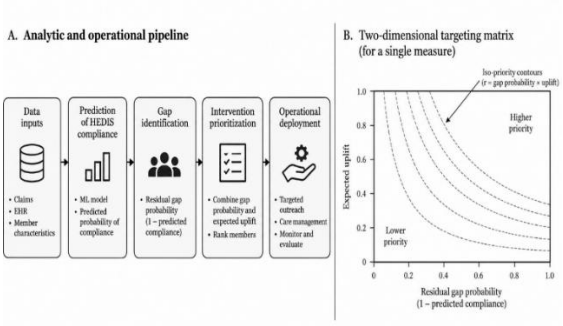


Figure 1 Framework for converting predicted HEDIS measure compliance into intervention priority

7. Validation, calibration and reporting

The parameter most frequently cited is the area under the receiver operating characteristic curve, or discrimination, and this is the least useful for this application. The numerically accurate predicted probabilities are essential to prioritization, since the priority score is the product of probabilities from different models, and is compared across measures. Those comparisons are skewed even if the ranking within a measure is maintained when scores are miscalibrated. Calibration should thus be evaluated as the first criterion based on calibration plots with flexible smoothers and calibration slope and intercept, and re-evaluated after each recalibration or resampling step [16,17]. Precision-recall summaries are more faithful than the receiver operating characteristic curve [18] when the prevalence of the gap is low, and they provide a description of the achievable yield. Thus, decision-curve analysis or similar, in terms of expected gaps closed per thousand contacts compared to current practice, is more informative to plan leadership than any global discrimination statistic [19]. Validations should be temporal (from one measurement year to a subsequent one) and if a plan is used in several markets, geographic (from one market to another market) as provider supply and documentation practice varies. Reporting should be done according to TRIPOD statement and risk of bias should be appraised using PROBAST when models are described in the peer-reviewed literature [20-22]. The evidence from the systematic-review is similar in other related fields such as prediction of hospital readmission risk [23]

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5776-5785
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0750

where most published prediction models are at high risk of bias, and the description of the models is incomplete, making it impossible to reproduce or reuse them. Quality programs that have their own internally developed models should be held to the same standard: If the model cannot be independently recreated, then it cannot be independently audited when the performance begins to wobble.

8. Equity, Fairness and unintended consequences

Prediction models that are developed using routinely collected data also reflect the inequities of the delivery system that generates the data. A widely used population health algorithm that used cost as a proxy for illness systematically underestimated the need of black patients, and a gap-closure model that predicts the probability of compliance and then deprioritizes members whose probability of compliance is judged to be low will withdraw resources from exactly those members whose access barriers make it unlikely that they will comply [24]. Here, efficiency and equity part ways, and it is not a modeling artifact but rather a property of the objective. There are 3 recommended safeguards. Firstly, stratification by dual eligibility, low-income subsidy status, race and ethnicity (when reliably recorded), and disability status and area deprivation should be reported in addition to the overall group; equal discrimination and unequal calibration will lead to unequal allocation. Second, the uplift framing should be intentional, as a member routed with structural barriers to case management will keep the attention of the system, but an appropriate intervention will be allocated. Third, capacity floors for high social-risk subgroups are a reasonable constraint to impose on an optimizer, and they are in the direction of the policy of the program, as CMS has shifted to rewarding performance among enrollees with social risk factors through a health equity index that replaced the previous reward factor [4]. Even when the measured rate is not adjusted, evidence of association of neighborhood disadvantage and social risk with worse measured chronic disease management suggests adjustment or stratification in internal management reporting [10,25].

9. Implementation considerations

There are a number of practical issues that will make the difference in whether a technically sound model alters measured performance. Scoring cadence should align with the operational calendar: monthly lists of refreshed scores with censored features are more useful than an annual score; scores refreshed late in the year should be downweighted if the window to close them has been effectively closed. Capacity should be explicitly considered as a constraint, as a ranked list is only meaningful in the context of the number of contacts the plan can make. There should be separation between provider-facing and member-facing outputs to support different decisions: Provider-level gap reports should be based on stable attribution and small-panel shrinkage, and member-level outreach should be based on channel preference and reachability. Evaluation should be planned prior to deployment. Pre-Post comparisons overestimate the program effect due to an increase in gap closure throughout the year, which is not the result of outreach. Even a small holdout or staggered rollout offers a concurrent counterfactual, creates variation for estimating uplift in future cycles, and is generally defensible if capacity isn't available to contact everyone in either case. Lastly, distribution shift should be performed on a regular basis. The relationship between features and labels can change due to benefit changes, network changes, measure specification revisions and changes of the completeness of supplemental data and a model that is not monitored will degrade silently [13].

10. Limitations

This review is narrative and does not attempt to be exhaustive in terms of the primary studies searched for; therefore, specific quantitative results regarding model performance in this domain should be looked for in the primary studies. The framework has not yet been tested in a prospective trial, and the priority score listed assumes that measure-level uplift estimates and costs are available, which will not be the case for most plans initially, and sequential adoption, starting with the prediction of a residual gap and then adding in the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5776-5785
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0750

uplift estimates and costs as the outreach variation is accumulated, is the realistic path. Program parameters (such as measure inventories, weights, cut-point methodology, equity adjustments) also vary from year to year and implementation requires re-specification to the current CMS and NCQA documentation.

Conclusion

Predictive modeling can enhance HEDIS performance in the Medicare Advantage Star Ratings program, but only if the amount of data modeled matches what the decision is. Plans should anticipate the gap at year-end that remains after features censored at the deployment date, estimate responsiveness to intervention independent of risk, incorporate both responsiveness and measure weight and feasibility into an explicit priority score, and validate calibration within social-risk strata, not just based on aggregate discrimination. Such a framework turns a descriptive gap list into a resource-allocation instrument, and can be audited by quality, compliance and equity stakeholders alike, when evaluated against a concurrent counterfactual and monitored for drift.

References

[1]. Centers for Medicare & Medicaid Services. Medicare 2025 Part C & D Star Ratings technical notes. Baltimore, MD: CMS; 2024.
[2]. Reid RO, Deb P, Howell BL, Shrank WH. Association between Medicare Advantage plan star ratings and enrollment. JAMA. 2013;309(3):267–274.
[3]. Newhouse JP, McGuire TG. How successful is Medicare Advantage? Milbank Q. 2014;92(2):351–394.
[4]. Centers for Medicare & Medicaid Services. Medicare program; contract year 2024 policy and technical changes to the Medicare Advantage and Medicare Prescription Drug Benefit programs. Final rule. Fed Regist. 2023;88:22120–22345.
[5]. Xu P, Burgess JF Jr, Cabral H, Soria-Saucedo R, Kazis LE. Relationships between Medicare Advantage contract characteristics and quality-of-care ratings: an observational analysis of Medicare Advantage star ratings. Ann Intern Med. 2015;162(5):353–358.
[6]. National Committee for Quality Assurance. HEDIS measures and technical resources. Washington, DC: NCQA. Available from: https://www.ncqa.org/hedis/
[7]. Ryan AM, Krinsky S, Kontopantelis E, Doran T. Long-term evidence for the effect of pay-for-performance in primary care. Lancet. 2016;388(10041):268–274.
[8]. Jacobson Vann JC, Jacobson RM, Coyne-Beasley T, Asafu-Adjei JK, Szilagyi PG. Patient reminder and recall interventions to improve immunization rates. Cochrane Database Syst Rev. 2018;1:CD003941.
[9]. Ash AS, Mick EO, Ellis RP, Kiefe CI, Allison JJ, Clark MA. Social determinants of health in managed care payment formulas. JAMA Intern Med. 2017;177(10):1424–1430.
[10]. Bhavsar NA, Gao A, Phelan M, Pagidipati NJ, Goldstein BA. Value of neighborhood socioeconomic status in predicting risk of outcomes in studies that use electronic health record data. JAMA Netw Open. 2018;1(5):e182716.
[11]. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441.
[12]. 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. New York: ACM; 2016. p. 785–794.
[13]. Rose S. Machine learning for prediction in electronic health data. JAMA Netw Open. 2018;1(4):e181404.
[14]. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–1358.
[15]. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30. 2017. p. 4765–4774.
[16]. Steyerberg EW, Vickers AJ, Cook NR, et

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5776-5785
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0750

al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–138.

[17]. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.

[18]. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.

[19]. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–574.

[20]. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55–63.

[21]. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1–W73.

[22]. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–58.

[23]. Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: a systematic review. JAMA. 2011;306(15):1688–1698.

[24]. 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):447–453.

[25]. Durfey SNM, Kind AJH, Buckingham WR, DuGoff EH, Trivedi AN. Neighborhood disadvantage and chronic disease management. Health Serv Res. 2019;54(Suppl 1):206–216.