

Machine Learning Versus Conventional Benchmarks for Predicting Next-Year Medicare Allowed Amounts

Tarun Siga

Independent Researcher, Kansas, USA; Email: tarunsiga7@gmail.com

*Corresponding Author: tarunsiga7@gmail.com

Abstract—The Medicare Physician Fee Schedule provides an administrative reference for physician payment, but fee-schedule values and historical averages are retrospective and may not capture nonlinear, geographic, specialty-specific, setting-specific, and utilization-related changes that influence future observed allowed amounts. This study evaluates whether machine-learning and conventional models can predict next-year service-weighted average Medicare allowed amounts for high-volume evaluation-and-management services more accurately than predetermined fee-schedule and historical benchmarks. Publicly available Centers for Medicare & Medicaid Services (CMS) Medicare Physician and Other Practitioners data and Physician Fee Schedule files from 2018–2024 were used. The analysis was restricted to participating individual physician billing for HCPCS 99213 and 99214 across the 50 U.S. states and the District of Columbia, separated by facility and non-facility setting. Groups required at least 100 services and five unique clinicians. Models were developed on earlier years, selected using 2023 data, and evaluated on a sealed 2024 test set. Histogram gradient boosting (HGB), random forest, ridge regression, pooled-history, recent-year-history, and fee-schedule-reference approaches were compared using service-weighted mean absolute error (WMAE). The final 2024 cohort contained 6,829 eligible groups representing 117,969,671 services. HGB achieved the lowest WMAE at \$0.829 per service, compared with random forest at \$1.173, ridge regression at \$1.650, pooled history at \$1.884, recent-year history at \$1.907, and the fee-schedule reference at \$4.803. Relative to ridge regression, HGB reduced WMAE by 49.7%, with a state-block bootstrap difference of -\$0.821 per service (95% CI: -\$1.172 to -\$0.563). HGB improved WMAE in 49 of 51 jurisdictions, with at least 5% improvement in 47. These findings support temporally sealed, group-level machine-learning forecasting of Medicare allowed amounts for the evaluated services while not establishing claims-level accuracy, causal effects, or future official CMS payment decisions.

Keywords—Medicare allowed amount; Physician Fee Schedule; machine learning; histogram gradient boosting; reimbursement forecasting; HCPCS; predictive modeling; weighted mean absolute error; temporal validation

Received: 16-06-2026; Revised: 26-08-2026; Accepted: 24-09-2026; Published: 30-09-2026

DOI: 10.22362/ijcert/2026/v13/i3/v13i3004

1 Introduction

Medicare physician services are reimbursed under the Physician Fee Schedule (PFS), in which work, practice-expense, and malpractice relative value units are converted to dollar payments after geographic adjustment and application of a national conversion factor. National health expenditure projections continue to show the scale and policy importance of Medicare and physician-service spending [1]. However, the observed Medicare allowed amount for a service group is not identical to a single administrative schedule value and can vary with procedure mix, specialty, geography, setting, and utilization. Recent Medicare reimbursement analyses also document temporal and geographic variation in payment patterns [2].

Historical averages and fee-schedule references are straightforward and interpretable, but they are primarily backward-looking. They are not designed to model nonlinear changes among code, specialty, state, setting, and time. In broader U.S. health-expenditure forecasting, machine-learning and statistical forecasting combinations have been investigated as flexible alternatives to purely conventional models [3], while machine-learning studies of health expenditure have shown that nonlinear methods can capture relationships that may be

missed by simpler specifications [4].

A central methodological requirement is to distinguish exploratory model development from confirmatory future-year evaluation. Preregistration and protocol prespecification are intended to reduce undisclosed analytical flexibility and strengthen interpretability of confirmatory findings [5, 6]. This consideration is particularly relevant to temporal prediction, because access to the eventual test-year outcome during model tuning can introduce leakage and yield optimistic performance estimates.

The study therefore addresses a specific predictive question: whether frozen models can forecast the next year's group-level Medicare allowed amounts for HCPCS 99213 and 99214 more accurately than prespecified administrative and historical comparators. The primary objective is to compare machine-learning models with conventional benchmarks on a sealed 2024 test year. Secondary objectives are to compare HGB, random forest, and ridge regression, evaluate performance across facility and non-facility settings, examine geographic heterogeneity across the 50 states and the District of Columbia, and quantify uncertainty through state-block bootstrap intervals.

The confirmatory hypothesis was specified before the sealed 2024 outcomes were evaluated. HGB was compared with

ridge regression, and success required both a relative WMAE reduction of at least 5% and a paired 95% confidence interval for the HGB-minus-ridge WMAE difference that excluded zero in the favorable direction. This design separates model selection from final evaluation and limits interpretation to the prespecified group-level prediction task.

2 Related Work

Medicare physician payment is an administered system rather than a market-price mechanism. Policy analyses have described structural limitations in the PFS and the interaction between fee-for-service valuation and value-based payment [7]. Statutory budget-neutrality requirements and shifts in service volume can also redistribute payments among provider types and specialties [8]. Consequently, observed allowed amounts reflect policy, geographic adjustment, provider composition, and utilization rather than a single unconstrained market value.

2.1 Payment Variation and Predictive Reimbursement Modeling

Payment variation is particularly relevant when technology, service delivery, or procedure mix changes over time. Medicare's approach to payment for artificial-intelligence-enabled services illustrates how evolving clinical technology can challenge existing fee-for-service structures [9]. A recent Medicare reimbursement preprint proposed a machine-learning pipeline for predicting changes across multiple fee schedules [10]; because it remains a preprint, it is treated here as supporting context rather than definitive peer-reviewed evidence.

Predictive modeling has also been applied to diagnostic and billing-code prediction from clinical documentation [11] and to multimodal financial-risk modeling for bundled payments in spinal surgery [12]. These studies demonstrate that structured and unstructured healthcare data can support payment-related predictive tasks, although their outcomes and units of analysis differ from Medicare group-level allowed amounts.

2.2 Machine Learning Versus Conventional Statistical Models

Comparative claims-based research has shown that tree ensembles can outperform conventional models in nonlinear prediction settings. Langenberger *et al.* compared random forest, gradient boosting, neural-network, and logistic-regression approaches for predicting future high-cost patients [13]. Related work has used electronic health-record behavior and administrative variables to predict physician departure, further illustrating the value of nonlinear interactions in operational health data [14]. These findings motivate the use of HGB and random forest as nonlinear comparators while retaining ridge regression as a regularized additive benchmark.

2.3 Temporal Validation and Model Drift

Future-year prediction requires evaluation under temporal change. Performance drift has been documented in machine-learning models for cardiac-surgery risk prediction, where temporal shifts affected model performance [15]. Longitudinal evaluation across hospitals likewise shows that clinical prediction models can experience changes in calibration and utility as data distributions evolve [16]. Although the present outcome is administrative rather than clinical, the underlying methodological principle is the same: a model intended for future-year use should be evaluated on data that occur after model development and selection.

2.4 Research Gap

Existing work provides strong foundations for Medicare payment policy, health-cost prediction, and temporal validation, but the source literature does not establish a same-cohort comparison of fee-schedule, historical, regularized-linear, random-forest, and histogram-gradient-boosting approaches for next-year group-level Medicare allowed amounts under a sealed national test year. The present study addresses that methodological gap using a single eligibility definition, common outcome metric, prespecified comparators, and temporally separated final evaluation.

3 Materials and Methods

The study used a multiyear retrospective design with temporally separated model development, selection, and final validation. Public CMS Medicare Physician and Other Practitioners data were combined with Physician Fee Schedule reference information for 2018–2024. The design follows general recommendations that prediction modeling clearly separate development from validation and report performance under conditions that approximate future use [17].

3.1 Data Sources and Cohort Construction

The outcome source was the CMS Medicare Physician and Other Practitioners Public Use File. Available fields included National Provider Identifier, state, ZIP code, provider specialty, HCPCS code, setting, service count, and average Medicare allowed amount. The administrative reference source was the quarterly PFS relative-value-unit release, with ZIP-to-locality information used to associate records with applicable geographic references. Open healthcare data have previously been used to develop interpretable forecasting models for healthcare costs [18].

The analysis was restricted to HCPCS 99213 and 99214 billed by participating individual physicians in the 50 states and the District of Columbia. Records failing specialty, participation, ZIP/locality, rate-state, or geographic checks were excluded. Remaining records were aggregated into Year $\times$ HCPCS $\times$ Specialty $\times$ State $\times$ Setting groups. A group was eligible only when it represented at least 100 services and at least five distinct billing NPIs. Minimum-count filtering reduces instability in sparse administrative groups and is consistent with the general need to control low-support observations in claims modeling [13, 19].

The same grouping rules were applied to development, selection, and final test years. For 2024, the initial retrieval contained 994,488 provider-service rows representing approximately 170.1 million services. After record-level screening, 603,168 rows representing about 118.5 million services remained for aggregation. A total of 9,556 candidate groups were formed, of which 6,829 met both eligibility thresholds and represented 117,969,671 services. Table 1 summarizes this construction before the final cohort is used in model evaluation.

The cohort construction yields four HCPCS-by-setting strata across 51 jurisdictions. All confirmatory comparisons reported below use this same eligible population so that differences in WMAE reflect prediction performance rather than changes in sample composition.

Table 1. 2024 confirmatory-cohort construction and exclusions.

Stage	Rows/groups	Services (approx.)
Initial 2024 retrieval	994,488 rows	170,077,047
After specialty, participation, ZIP/locality, rate-match, and geographic checks	603,168 rows	118,541,272
Candidate groups formed before thresholds	9,556 groups	—
Eligible groups in final confirmatory cohort	6,829 groups	117,969,671
Excluded below 100-service or 5-NPI threshold	2,727 groups	—

3.2 Outcome, Predictors, and Prediction Target

The prediction target was the service-weighted group mean Medicare allowed amount. Predictors included year, HCPCS code, specialty, state, setting, and the applicable fee-schedule reference. The models predicted the relationship between observed allowed amount and the administrative reference, with the future-year reference incorporated when generating the next-year estimate. Training weights used service counts and reference information, whereas final evaluation weights were based on service counts.

3.3 Predictive Methods and Comparators

Six prespecified approaches were evaluated: HGB, random forest, ridge regression, pooled-history prediction, recent-year-history prediction, and direct fee-schedule reference. Historical rules used hierarchical fallback logic when fine-grained groups did not meet support requirements. Ridge regression used one-hot encoded categorical variables with $\alpha = 10$. Random forest used 350 trees, a minimum leaf size of 5, and 80% feature subsampling. HGB used 300 iterations, learning rate 0.05, 31 leaves, minimum leaf size 20, and L2 regularization. Tree ensembles were included because they can represent non-linear interactions that are difficult to express with additive linear models [13].

3.4 Temporal Development, Selection, and Protocol Freeze

Expanding-year development evaluations were conducted through 2022, and 2023 was used for final model selection. The 2024 outcome was kept sealed until eligibility rules, predictors, comparators, metrics, bootstrap strategy, decision threshold, and analysis code were frozen. This chronology follows the broader principle that model evaluation should reflect data not used during tuning or model choice [17, 20].

3.5 Evaluation Metrics and Statistical Uncertainty

The primary metric was service-weighted mean absolute error. Equation 1 defines the metric used for all confirmatory comparisons:

$$\text{WMAE} = \frac{\sum_g s_g |y_g - \hat{y}_g|}{\sum_g s_g}, \quad (1)$$

where g indexes eligible groups, s_g is service count, y_g is the observed group mean allowed amount, and $\hat{y}_g$ is the predicted amount. Secondary metrics were signed error and unweighted group MAE. The confirmatory contrast was HGB WMAE minus ridge WMAE, accompanied by a paired 95% interval from 2,000 state-block bootstrap resamples. Reporting interval estimates in addition to point performance follows recommendations for evaluation of prediction models [20].

The relative WMAE reduction used in the confirmatory

decision is defined in Equation 2:

$$R = 100 \times \frac{\text{WMAE}_{\text{ridge}} - \text{WMAE}_{\text{HGB}}}{\text{WMAE}_{\text{ridge}}}. \quad (2)$$

This percentage was interpreted together with the absolute bootstrap difference; neither measure was treated as evidence of a causal payment effect.

3.6 Subgroup, Geographic, and Reproducibility Analyses

Prespecified secondary analyses examined all four HCPCS-by-setting strata and all 51 jurisdictions. Jurisdiction-level summaries reported the number of locations with any HGB improvement, the number reaching the 5% threshold, the median reduction, and the extreme jurisdictions. These analyses were descriptive rather than independent confirmatory tests.

The reproducibility audit retained manifests, hashes, cohort logs, exclusions, metrics, and analysis records. The source manuscript reports review of 1,435 development pages, 250 final-year pages, 35 rate files, 3,488 checks, and 56 tracked files. Because the study used publicly available administrative data without patient identifiers, no new patient recruitment or intervention was performed; institutional requirements for formal ethics review or exemption nevertheless remain an author/institutional responsibility.

4 Results

The sealed 2024 cohort contained 6,829 eligible groups and 117,969,671 services. Results are reported first for the prespecified overall comparison, followed by HCPCS-setting strata and geographic heterogeneity. The presentation retains the same cohort definition across all methods.

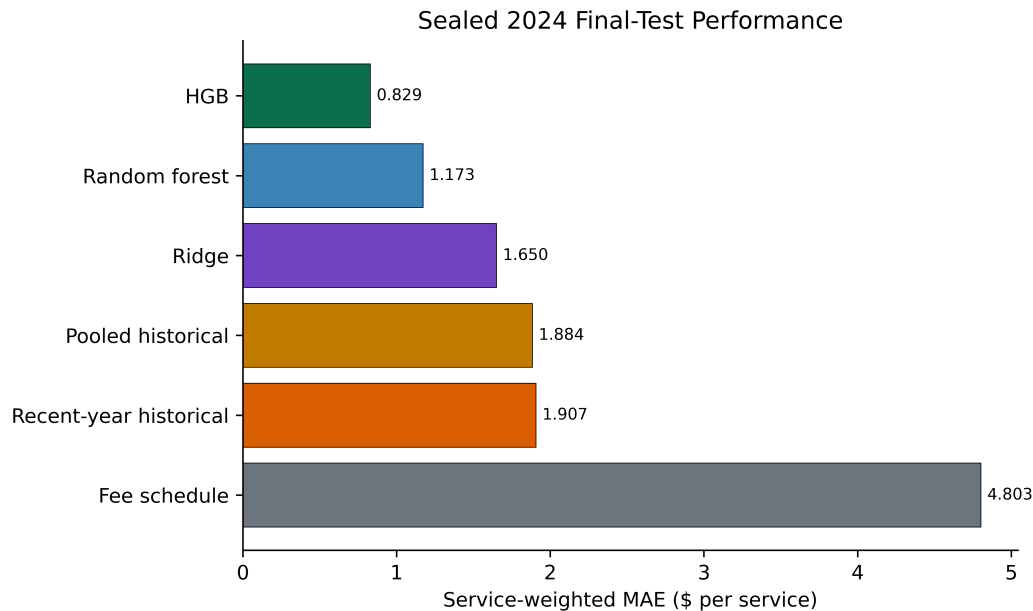
4.1 Overall Sealed 2024 Performance

HGB achieved the lowest WMAE of all six prespecified methods at \$0.829 per service. Random forest followed at \$1.173, ridge regression at \$1.650, pooled history at \$1.884, recent-year history at \$1.907, and the fee-schedule reference at \$4.803. Relative to ridge regression, HGB reduced WMAE by 49.7%, corresponding to an absolute difference of -\$0.821 per service. The state-block bootstrap 95% interval was -\$1.172 to -\$0.563, satisfying both prespecified confirmatory criteria. Table 2 reports the full metric set before Figure 1 visualizes the primary error comparison.

The error ordering in Table 2 shows that both tree ensembles outperformed the regularized additive model, while the two historical rules were slightly worse than ridge regression. The fee-schedule reference had the largest WMAE and a positive signed error of \$4.791 per service. Figure 1 displays the same WMAE values without introducing additional derived estimates.

Table 2. Sealed 2024 performance across all six prespecified methods.

Method	WMAE (\$/svc)	Signed error (\$/svc)	Unweighted group MAE (\$)	Reduction vs. ridge
Histogram gradient boosting	0.829	+0.083	1.630	49.7%
Random forest	1.173	+0.139	1.734	28.9%
Ridge regression	1.650	+0.649	2.365	–
Pooled historical	1.884	+0.694	2.076	-14.2%
Recent-year historical	1.907	+0.054	2.044	-15.6%
Fee-schedule reference	4.803	+4.791	3.926	-191.1%

**Figure 1.** Service-weighted MAE by method in the sealed 2024 final test. Values are reproduced from the prespecified overall comparison.

The overall result supports the prespecified predictive comparison but does not by itself establish uniform performance across code-setting combinations or jurisdictions. Those secondary checks are therefore reported separately below.

4.2 Performance by HCPCS Code and Setting

HGB had lower WMAE than ridge regression in each of the four predefined HCPCS-by-setting strata. The reduction ranged from 38.6% to 53.0%, and HGB WMAE remained below \$1.00 per service in every stratum. Table 3 provides the stratum-level values before Figure 2 compares fee schedule, ridge, and HGB graphically.

The largest absolute service volume occurred in the 99214 non-facility stratum, where HGB WMAE was \$0.937 compared with \$1.993 for ridge and \$5.870 for the fee-schedule reference. Figure 2 shows that the relative ordering between HGB and ridge was consistent across all four strata.

The stratum analysis indicates that the overall result was not driven by a single HCPCS-setting combination. Geographic heterogeneity remained, however, and was examined at the jurisdiction level.

4.3 Geographic Heterogeneity

Across the 50 states and the District of Columbia, HGB had lower WMAE than ridge in 49 of 51 jurisdictions. Forty-seven jurisdictions achieved at least the prespecified 5% reduction. The median jurisdiction-level reduction was 48.4%, with the largest reported reduction in Colorado at 77.6%. Florida and

New Mexico did not improve, with reported reductions of -22.6% and -10.3%, respectively. Table 4 summarizes the geographic results before Figure 3 presents the jurisdiction counts by outcome category.

The jurisdiction counts provide a compact view of geographic consistency. Figure 3 groups the 51 jurisdictions into those reaching at least 5% improvement, those improving by less than 5%, and those showing no improvement.

The geographic analysis supports broad, but not universal, improvement. The two jurisdictions without improvement require further investigation of specialty composition, group eligibility, locality mapping, and temporal change before any operational use is considered.

4.4 Reproducibility Audit

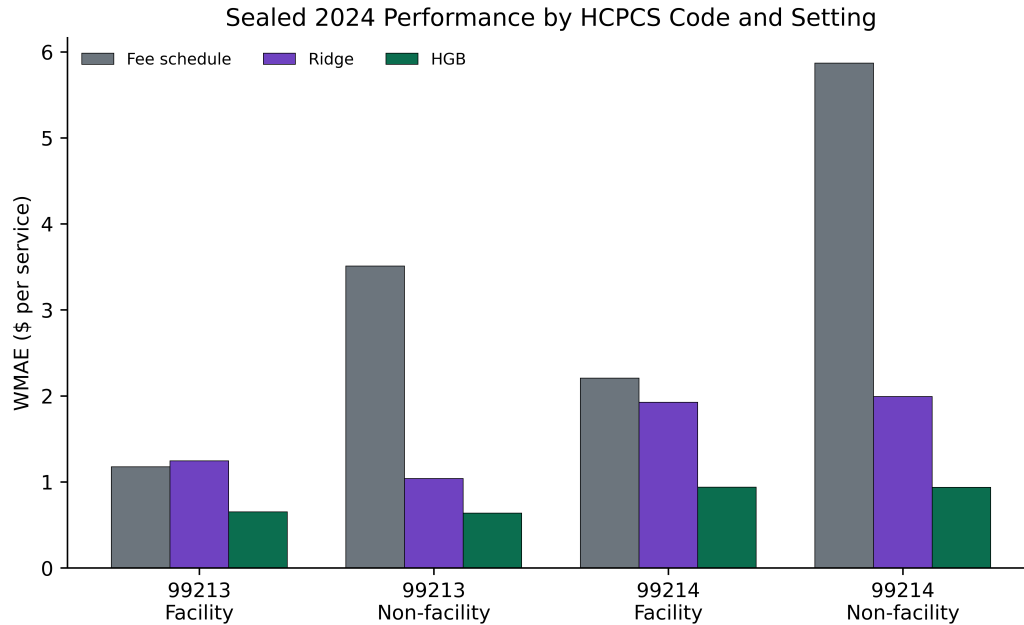
The reported audit compared the frozen analytical protocol with the executed analysis and retained source manifests, hashes, cohort logs, and result checks. The source manuscript reports 1,435 development pages, 250 final-year pages, 35 rate files, 3,488 checks, and 56 tracked files. No discrepancy was reported between the hashed protocol and the final analysis. These checks support internal reproducibility of the reported endpoints but do not replace independent external replication.

5 Discussion

The principal finding is that a frozen HGB model predicted 2024 group-level Medicare allowed amounts more accurately than the prespecified conventional and administrative com-

Table 3. Sealed 2024 performance by HCPCS code and setting.

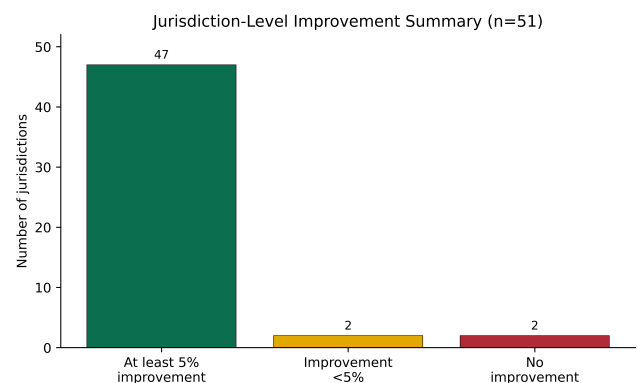
HCPCS	Setting	Groups	Services	Fee (\$)	Ridge (\$)	HGB (\$) / Reduction
99213	Facility	1,258	2,430,512	1.177	1.246	0.653 / 47.6%
99213	Non-facility	2,014	40,111,808	3.511	1.039	0.638 / 38.6%
99214	Facility	1,457	5,427,813	2.207	1.927	0.940 / 51.2%
99214	Non-facility	2,100	69,999,538	5.870	1.993	0.937 / 53.0%

**Figure 2.** Sealed 2024 WMAE by HCPCS code and facility setting for the fee-schedule reference, ridge regression, and histogram gradient boosting.**Table 4.** Confirmatory and geographic summary.

Measure	Result
Eligible jurisdictions	51 (50 states + DC)
Jurisdictions with HGB improvement	49 of 51
Jurisdictions with at least 5% improvement	47 of 51
Median jurisdiction reduction	48.4%
Largest reduction	Colorado, 77.6%
No improvement	Florida, -22.6%; New Mexico, -10.3%
Confirmatory absolute difference	-\$0.821 per service
State-block 95% interval	-\$1.172 to -\$0.563 per service
Confirmatory decision	Threshold met; interval excluded zero

parators for HCPCS 99213 and 99214. The 49.7% WMAE reduction relative to ridge regression was accompanied by a bootstrap interval that excluded zero, and the same direction of improvement was observed in all four HCPCS-setting strata and 49 of 51 jurisdictions.

The result is consistent with broader evidence that tree ensembles can represent nonlinear interactions in administrative and claims data [13]. In this application, specialty, geography, setting, reference amount, and time can interact in ways that are difficult to capture with a purely additive model. The find-

**Figure 3.** Jurisdiction-level summary of HGB improvement relative to ridge regression in the sealed 2024 test.

ing should therefore be interpreted as evidence about predictive structure in this specific grouped dataset, not as evidence that HGB is universally preferable for Medicare forecasting.

The sealed temporal design is an important methodological strength. Guidance on prediction-model development and evaluation emphasizes separation of development and validation data and careful reporting of external or temporally separated performance [17, 20]. The present protocol additionally froze cohort rules, predictors, comparators, metrics, bootstrap strategy, and decision criteria before opening the 2024 outcome. This reduces, but does not eliminate, risks of analytical overfitting.

5.1 Relation to Prior Healthcare-Cost Modeling

Machine-learning cost prediction has increasingly incorporated interpretability, uncertainty, and fairness. Explainable XGBoost approaches have been proposed for healthcare cost prediction using SHAP and LIME [21]. Open-data studies have likewise developed interpretable models for forecasting healthcare costs [18], and recent insurance-cost work has combined stacked ensembles, conformal intervals, SHAP explanations, and fairness auditing [22]. The present study differs by focusing on next-year group-level Medicare allowed amounts with a sealed administrative test year rather than individual insurance expenditure.

5.2 Strengths

Four features strengthen the evidence. First, the final year was temporally sealed rather than randomly mixed into development data. Second, all six comparators were evaluated on the same eligible cohort and primary metric. Third, the confirmatory criterion combined a relative threshold with a paired state-block interval. Fourth, subgroup and jurisdiction analyses made heterogeneity visible rather than relying only on a national aggregate.

5.3 Limitations

The findings are restricted to two evaluation-and-management codes and should not be generalized to specialty procedures, facility claims, or other billing environments without additional validation. The lack of improvement in Florida and New Mexico indicates that national averages can conceal meaningful geographic variation. Equity and subgroup performance should be examined explicitly before deployment; the HEAL framework provides one example of a structured approach to health-AI equity assessment [23].

The design also does not support causal inference. A lower forecast error does not identify why payment patterns changed, nor does it establish that the model predicts future CMS policy decisions. The direct fee-schedule comparator should be interpreted as an administrative reference rather than as a model designed to forecast observed group means. One related Medicare machine-learning source used in the background literature remains a preprint and has not yet undergone journal peer review [10]. In addition, group aggregation can mask clinician-level or claim-level heterogeneity.

5.4 Implications and Future Work

For payers, analysts, and health-system planners, a temporally validated machine-learning forecast can serve as an additional benchmark alongside administrative fee schedules and historical averages. It should not replace the official payment system. Operational use would require ongoing monitoring for model shift, periodic recalibration, transparent version control, and jurisdiction-specific auditing because temporal degradation has been documented in other healthcare prediction settings [15, 16].

Future studies should extend the same sealed-year protocol to additional HCPCS families and specialties, test multi-year stability, assess calibration and interval coverage, and incorporate interpretable feature-attribution analyses. Conformal prediction could be used to quantify forecast uncertainty, while SHAP-based diagnostics and fairness audits could identify groups for which accuracy is systematically lower [22]. In-

dependent replication using the same eligibility rules would further clarify external validity.

6 Conclusion

This study compared six prespecified approaches for predicting next-year group-level Medicare allowed amounts for HCPCS 99213 and 99214. In the sealed 2024 test, histogram gradient boosting achieved a service-weighted mean absolute error of \$0.829 per service, compared with \$1.650 for ridge regression and \$4.803 for the direct fee-schedule reference. The HGB-versus-ridge difference was -\$0.821 per service, corresponding to a 49.7% reduction, with a state-block 95% confidence interval from -\$1.172 to -\$0.563. Improvement was observed in all four HCPCS-by-setting strata and in 49 of 51 jurisdictions, although Florida and New Mexico did not improve. The evidence therefore supports HGB as a useful forecasting approach for the evaluated group-level task under the reported temporal protocol. It does not establish claims-level accuracy, causal payment effects, performance for other procedures, or prediction of future official CMS fee-schedule decisions.

Declarations

Acknowledgements. No specific acknowledgements were provided in the source manuscript.

Funding. The source manuscript did not report a specific external funding source. The author should confirm this statement before final publication.

Author Contributions. Tarun Siga is the sole listed author and is responsible for the study conception, data preparation, methodology, analysis, interpretation, visualization, and manuscript preparation, subject to the author's final confirmation before publication.

Conflict of Interest / Competing Interests. No conflict-of-interest statement was provided in the source manuscript. The author should confirm whether any competing financial or non-financial interests exist before publication.

Data Availability. The study uses publicly available Centers for Medicare & Medicaid Services Medicare Physician and Other Practitioners data and Physician Fee Schedule files for 2018–2024. The exact extraction manifests and processed analytical files should be supplied or archived by the author if required by the journal.

Code Availability. The source manuscript reports a reproducibility process based on manifests, hashes, cohort logs, exclusions, metrics, and audits. A public code repository was not identified in the submitted manuscript; availability should be confirmed by the author.

Ethics Approval. The reported analysis used publicly available administrative data without patient identifiers and involved no new participant recruitment or intervention. Any requirement for institutional ethics review, exemption, or formal determination should be confirmed by the author and the responsible institution before publication.

Informed Consent. No new participants were recruited for the reported analysis. Applicability of formal consent requirements should be confirmed in accordance with institutional policy.

Consent for Publication. The manuscript does not report newly collected identifiable participant information. The author should confirm that no additional publication consent is required.

Generative AI and AI-Assisted Technologies. The final disclosure of any generative-AI or AI-assisted use must be confirmed by the author in accordance with IJCERT policy. Scientific responsibility for the methodology, results, references, interpretations, and final manuscript remains with the author.

Additional Information / Supplementary Material. No supplementary material was included with the submitted source manuscript. Reproducibility files may be supplied separately if required.

References

- [1] J. A. Fiore, A. M. Sisko, J. A. Poisal, S. D. Smith, G. A. Cuckler, A. J. Madison, S. P. Keehan, K. E. Rennie, and N. J. Feehley, “National Health Expenditure Projections, 2025–34: Strong Utilization Growth Initially, Legislative Impacts Later,” *Health Affairs*, vol. 45, no. 7, pp. 722–733, 2026, doi: 10.1377/hlthaff.2026.00642.
- [2] A. Warriar, S. Iqbal, D. Bu, R. Bartholomew, and D. J. Lee, “Spatiotemporal Medicare Reimbursement Trends and Economic Clusters in Otolologic/Neurologic Surgery,” *The Laryngoscope*, vol. 136, no. 10, pp. 4444–4456, 2026, doi: 10.1002/lary.70652.
- [3] J. Wang, Z. Qin, J. Hsu, and B. Zhou, “A fusion of machine learning algorithms and traditional statistical forecasting models for analyzing American healthcare expenditure,” *Healthcare Analytics*, vol. 5, Art. no. 100312, 2024, doi: 10.1016/j.health.2024.100312.
- [4] N. Caravaggio, R. Lagravinese, and G. Resce, “The determinants of health expenditure: a machine learning approach,” *Empirical Economics*, vol. 70, Art. no. 30, 2026, doi: 10.1007/s00181-025-02854-6.
- [5] T. S. Dee, “The case for preregistering quasi-experimental program and policy evaluations,” *Evaluation Review*, vol. 49, no. 5, pp. 931–945, 2025, doi: 10.1177/0193841X251326738.
- [6] D. Lakens, C. Mesquida Caldentey, S. Rasti, and M. Ditroilo, “The benefits of preregistration and Registered Reports,” *Evidence-Based Toxicology*, vol. 2, no. 1, Art. no. 2376046, 2024, doi: 10.1080/2833373X.2024.2376046.
- [7] R. A. Berenson and K. J. Hayes, “The road to value can’t be paved with a broken Medicare Physician Fee Schedule,” *Health Affairs*, vol. 43, no. 7, pp. 950–958, 2024, doi: 10.1377/hlthaff.2024.00299.
- [8] E. W. Christensen, G. N. Nicola, E. Y. Rula, L. P. Nicola, and J. A. Hirsch, “Medicare Volume Growth and Shift in Payments From Physicians to Non-Physician Practitioners Under Statutory Budget Neutrality,” *Inquiry*, vol. 61, Art. no. 00469580241249076, 2024, doi: 10.1177/00469580241249076.
- [9] A. Zink, M. E. Chernew, and H. T. Neprash, “How Should Medicare Pay for Artificial Intelligence?” *JAMA Internal Medicine*, vol. 184, no. 8, pp. 863–864, 2024, doi: 10.1001/jamainternmed.2024.1648.
- [10] I. H. Patel, A. Leyva, and M. K. K. Niazi, “A Multi-Schedule Machine Learning Pipeline for Medicare Reimbursement Change Prediction and Operational Risk Stratification,” *medRxiv*, preprint, 2026, doi: 10.64898/2026.07.22.26358736.
- [11] A. Rajaram, M. Judd, and D. Barber, “Deep Learning Models to Predict Diagnostic and Billing Codes Following Visits to a Family Medicine Practice: Development and Validation Study,” *JMIR AI*, vol. 4, Art. no. e64279, 2025, doi: 10.2196/64279.
- [12] K. A. Mani, S. N. Goldman, T. Scharfenberger, V. Shankar, M. Bhatta, R. De la Garza Ramos, M. S. Fourman, and A. S. Eleswarapu, “Multimodal machine learning for risk-stratified bundled payments in spinal surgery,” *npj Digital Medicine*, vol. 8, Art. no. 512, 2025, doi: 10.1038/s41746-025-01915-5.
- [13] B. Langenberger, T. Schulte, and O. Groene, “The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data,” *PLOS ONE*, vol. 18, no. 1, Art. no. e0279540, 2023, doi: 10.1371/journal.pone.0279540.
- [14] K. Lopez, H. Li, H. Paek, B. Williams, B. Nath, E. R. Melnick, and A. J. Loza, “Predicting physician departure with machine learning on EHR use patterns: A longitudinal cohort from a large multi-specialty ambulatory practice,” *PLOS ONE*, vol. 18, no. 2, Art. no. e0280251, 2023, doi: 10.1371/journal.pone.0280251.
- [15] T. Dong *et al.*, “Performance Drift in Machine Learning Models for Cardiac Surgery Risk Prediction: Retrospective Analysis,” *JMIRx Med*, vol. 5, Art. no. e45973, 2024, doi: 10.2196/45973.
- [16] P. C. Silva *et al.*, “Longitudinal Model Shifts of Machine Learning-Based Clinical Risk Prediction Models: Evaluation Study of Multiple Use Cases Across Different Hospitals,” *Journal of Medical Internet Research*, vol. 26, Art. no. e51409, 2024, doi: 10.2196/51409.
- [17] O. Efthimiou, M. Seo, K. Chalkou, T. Debray, M. Egger, and G. Salanti, “Developing clinical prediction models: a step-by-step guide,” *BMJ*, vol. 386, Art. no. e078276, 2024, doi: 10.1136/bmj-2023-078276.
- [18] A. R. Rao, R. Jain, M. Singh, and R. Garg, “Predictive interpretable analytics models for forecasting healthcare costs using open healthcare data,” *Healthcare Analytics*, vol. 6, Art. no. 100351, 2024, doi: 10.1016/j.health.2024.100351.
- [19] E. Brati, A. Braimllari, and A. Gjeçi, “Machine Learning Applications for Predicting High-Cost Claims Using Insurance Data,” *Data*, vol. 10, no. 6, Art. no. 90, 2025, doi: 10.3390/data10060090.

- [20] G. S. Collins *et al.*, “Evaluation of clinical prediction models (part 1): from development to external validation,” *BMJ*, vol. 384, Art. no. e074819, 2024, doi: 10.1136/bmj-2023-074819.
- [21] G. R. Ramya, M. V. Krishna, S. T. Nori, K. P. Yeshwanth, and G. K. Jagannadh, “Interpretable Healthcare Cost Prediction Using Explainable XGBoost with SHAP and LIME,” in *Sixth Congress on Intelligent Systems*, pp. 47–61, 2026, doi: 10.1007/978-3-032-18135-0_5.
- [22] A. Al Ataby, F. Moutassem, and W. Al Nuaimy, “Insurance Cost Prediction with Stacked Ensembles, Conformal Intervals, SHAP Explainability, and Fairness Auditing,” *Engineering, Technology & Applied Science Research*, vol. 16, no. 4, pp. 38276–38286, 2026, doi: 10.48084/etasr.19923.
- [23] M. Schaekermann *et al.*, “Health equity assessment of machine learning performance (HEAL): a framework and dermatology AI model case study,” *eClinicalMedicine*, vol. 70, Art. no. 102479, 2024, doi: 10.1016/j.eclinm.2024.102479.