

Machine-Learning Cost Prediction Models for Identifying High-Burden Patients in U. S. Insurance Markets

Sarah Paul¹; Adebayo O. Sosanya²; Edward Akuffo Addo Dankwa³

¹University of North Carolina Charlotte, North Carolina, United States.

²San Francisco Bay University, Fremont, California, United States.

³Mississippi State University, Starkville, Mississippi, United States.

Publication Date: 2026/10/09

Abstract: Few U.S. patients account for most healthcare spending. This explains why they are easily identified by machine learning (ML). This review synthesizes empirical evidence on the performance of cost-prediction models, their application, and overall implications in the U.S. insurance markets. There are claims that data have built models capable of identifying high-cost patients with strong accuracy, where random forest and gradient boosting models attained AUCs of 0.878 to 0.883, including a large-scale U.S. study covering 48 million insured lives attaining an AUC of 0.912 for identifying patients who are likely to exceed \$250,000 in annual costs – a group that represents only 0.16% of members but accrues up to 9% of total spending. Findings show that insurers and health systems use these tools to target expensive care-management programs. Correspondingly, the most significant finding in the literature is not about accuracy, but what happens when cost replaces health need. Precisely, a landmark study showed that a widely used commercial algorithm systematically underestimated the extent of sickness of Black patients because it predicted future cost instead of future illness. Meanwhile, unequal access to care is an indication that Black patients historically generated lower costs for a similar level of sickness. Hence, correcting this single design choice is more than twice the share of Black patients flagged for extra care [from 17.7% to 46.5%]. Therefore, this study concludes that cost-prediction models are effective for conventional accuracy standards, but accuracy alone does not provide information about fairness, while the specific choice of what a model is trained to predict, cost against illness, can determine its actual function if it helps or harms the patients it is built to serve.

Keywords: Machine Learning, Cost-Prediction Models, High-Burden Patients, U.S. Insurance Markets.

How to Cite: Sarah Paul; Adebayo O. Sosanya; Edward Akuffo Addo Dankwa (2026) Machine-Learning Cost Prediction Models for Identifying High-Burden Patients in U. S. Insurance Markets. *International Journal of Innovative Science and Research Technology*, 11(10), 135-140. <https://doi.org/10.38124/ijisrt/26oct219>

I. INTRODUCTION

Healthcare spending in the United States is skewed (Jones & Dolsten, 2024). A small group of patients, usually those with chronic conditions managed by multiple doctors, represent a disproportionate share of U.S. annual health spending (Hajat, Siegal, & Adler-Waxman, 2021). Health-need patients, defined as the top 5% spenders in a year, are estimated to use up to half of total healthcare costs (Osawa et al., 2020). The curve also shows that in a 2019 analysis of claims data that cover 48 million insured Americans, high-cost claimants, that is, people with annual costs that exceed \$250,000, represent only 0.16% of the insured population, while accounting for 9% of the total healthcare spending (Maisog et al., 2019).

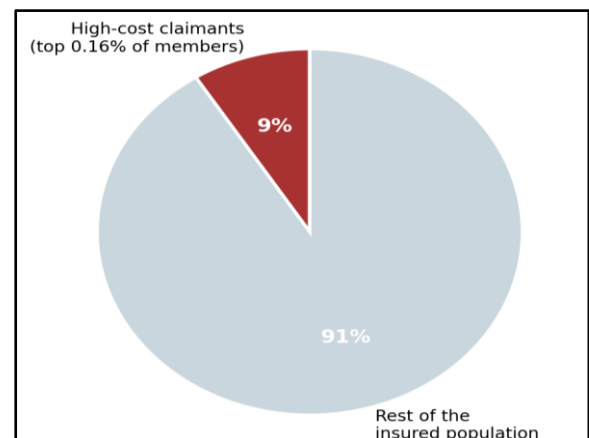


Fig 1 Share of Total U.S. Healthcare Spending for High-Cost Claimants and Other Insured Population [Adapted from Maisog et al., 2019].

Figure 1 above shows that a small sliver of patients drives approximately a tenth of all U.S. healthcare spending. The imbalance points to why insurers, employers, and health systems want to know and be able to predict who is likely to become a high-cost patient before their condition deteriorates and costs increase (Crowley et al., 2020; Chang, Mirvis, & Mahmood, 2026). Hence, in principle, if a plan can successfully identify a high-risk member early enough, it can enrol them in a care-management program, a chronic-disease coaching service, or a nurse check-in line. Consequently, this will improve their health while reducing downstream spending (Shockney, 2021). Machine learning (ML) has been identified as a viable tool for such prediction because of its inherent ability to find patterns across thousands of clinical variables and claims that would be ignored by a simple checklist or rule (Hutchins, 2025; Delanerolle et al., 2026). Based on recent evidence, this concentration of spending among a small proportion of patients is not limited to U.S. adult populations. A Belgian analysis of intensive care unit stays found similar skewed costs toward a subset of high-severity admissions (Bruyneel et al., 2023), while newer pediatric-focused studies in Tanzania and China report the same pattern of a small group of children making up disproportionate spending (Mwakinyali, Kateule, & Tunga, 2024; Zhang et al., 2025).

This review answers two questions:

- How accurate are the models currently in practice, and which approaches give the best performance?
- Does predicting cost imply predicting need, or can the two be distinguished without harming the patients?

II. DATA SOURCES AND METHOD

This review draws on peer-reviewed studies that built and tested machine learning (ML) models for healthcare cost or high-cost patient prediction using real clinical data or insurance claims, as well as notable fairness studies that examined the outcome of deploying the models at scale. Hence, this review is anchored on three distinct types of evidence:

- Model-comparison models that test diverse machine learning algorithms against each other on similar claims datasets to identify the ones that predict future high-cost

patients most accurately. A typical example is a German sickness-fund study which compares random forest, neural network, gradient boosting, and logistic regression models (Langenberger, Schulte, & Groene, 2023).

- Large-scale U.S. claims studies that build cost-prediction models on real-world insurance populations, including a study that used claims from nearly 50 million insured Americans to predict very high-cost claimants (Maisog et al., 2019), and a Texas Medicaid study, which covers more than 1.7 million adult enrollees (Yang, Delcher, Shenkman, & Ranka, 2018).
- The fairness and bias literature examining the outcome of a cost-prediction algorithm is deployed in a real healthcare system, centered on Obermeyer et al.'s 2019 Science study, auditing a commercial algorithm for managing the health of more than 100 million Americans.

Likewise, this review incorporates more recent methodological contributions that extend the evidence base, including a machine learning (ML) framework for predicting healthcare utilization and related risk factors (Rahman & Dua, 2025) as well as a channel-wise deep learning architecture for reducing prediction error and payment disparities among high-need patients with heterogenous clinical profiles (Morid & Sheng, 2025).

Throughout this paper, cost prediction refers to models trained to forecast a patient's future healthcare spending, and is distinguished from models trained to predict future illness or clinical need directly.

III. RESULTS

➤ Model Accuracy

Machine learning models can predict future high-cost patients (Osawa et al., 2020). Studies reviewed show consistently strong accuracy, with a fairly consistent ranking of methods. In the German study training random forest, artificial neural network, gradient boosting, and logistic regression models on three years of sickness-fund claims data, researchers found that two tree-based methods demonstrated the best performance: random forest reached an AUC of 0.883 and gradient boosting reached 0.878, which are collectively ahead of the neural network and logistic regression baseline (Langenberger et al., 2023).

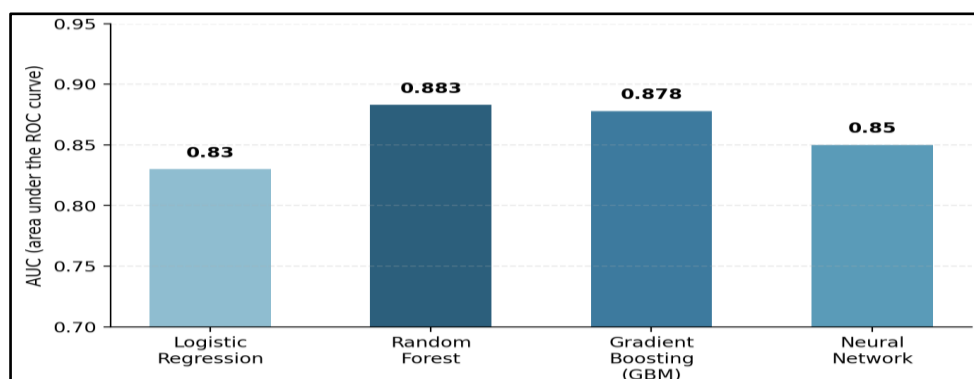


Fig 2 Reported Model Accuracy of Different Machine Learning Approaches for Predicting Future High-Cost Patients from Claims Data [Adapted from Langenberger et al., 2023].

As shown in Figure 2, at U.S. scale, performance sometimes holds up or improves. The study built specifically to identify very high-cost claimants, exceeding \$250,000 annually, reached an AUC of 0.912, which exceeded the highest previously published performance benchmark of 0.84 for such a task (Maisog et al., 2019). In a separate analysis of Texas Medicaid claims, covering 1.7 million adult enrollees, results showed that healthcare expenditures could be predicted with $R^2 > 0.7$, and prediction accuracy was strongest for the high-need, high-cost patients the models were designed to capture (Yang et al., 2018).

A common theme across these studies is that more clinical detail is helpful. According to the Japanese national screening study, combining clinical data, lab values such as HbA1c and LDL cholesterol, blood pressure, and survey responses on smoking and medication history, with insurance claims data produces a c-statistic of 0.84, outperforming models built on only claims data (Osawa et al., 2020). This shows that cost prediction improves when a model can get a signal of the patient's actual physiology, rather than just their billing history.

This pattern of strong tree-based and ensemble performance is recurrent in more recent studies beyond the U.S. and German cohorts. A 2025 study predicting high-need, high-cost pediatric hospitalizations in China found that ensemble methods were superior to simpler baselines (Zhang et al., 2025). Likewise, a machine learning model built to flag high-cost, high-need intensive care patients from available data within the first 24 hours of admission achieved good calibration and discrimination with a random forest approach (Nakano et al., 2025).

➤ Real Life Applications of the Models

Cost-prediction models designed for this purpose are not limited to academic activities. They feed certain operational decisions, involving patients' enrollment in a care-management program, a higher tier of case-management attention, or a proactive outreach call (Mo, 2026). This is the algorithm's function at the center of the fairness study, and the goal of most claims-based prediction studies reviewed in this paper. For example, the Texas Medicaid study frames its

prediction work around identifying high-need, high-cost patients for targeted intervention, on the premise that it is easier to change a patient's trajectory with a health plan when captured early enough (Yang et al., 2018).

➤ Fairness Failure

The most significant study in this literature on ML cost prediction for high-burden patients is not limited to accuracy. Obermeyer et al. (2019), from the examination of a widely used commercial algorithm which was deployed across the U.S. health system and influencing care decisions for over 100 million Americans using predicted future healthcare cost as a proxy for the extent of sickness of a patient, to decide which patient should be enrolled in a high-risk care-management program, compared the algorithm's risk scores against direct measures of patient health. This included the number of chronic conditions a patient had and other biomarkers of uncontrolled illness, for more than 43,000 White patients and 6,000 Black patients enrolled at the hospital system, which represented the source of the data.

The results showed that at any given risk score assigned by the algorithm, Black patients were considered to be sicker than White patients. The algorithm did not make a cost-prediction error. It was accurate at the task (Obermeyer et al., 2019). However, cost and illness were different for the two groups of patients. Due to longstanding, unequal access to care, Black patients with a level of illness historically produced lower healthcare costs than White patients with the same level of illness (Dickman et al., 2022). This is because they received less care for their sickness. A cost-predicting model learned to systematically underestimate how sick Black patients are (Obermeyer et al., 2019).

Meanwhile, the scale of the consequence was massive. According to researchers, this single design choice involved the use of cost instead of illness as the prediction target, which cut the proportion of Black patients identified for extra care by nearly 50% compared to what would have been identified by an unbiased algorithm. Hence, correcting the algorithm to predict health need directly, rather than cost, raised the proportion of Black patients identified for extra care management [See Figure 3].

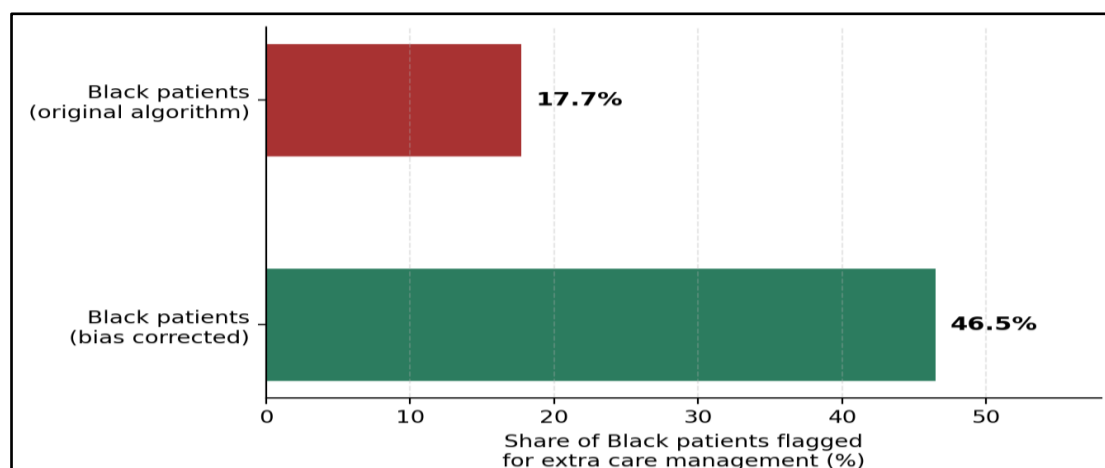


Fig 3 Share of Black Patients Identified for Extra Care Management – Before and After Correcting the Prediction Target of the Algorithm from Cost to Health Need

This finding is core to the broader machine learning field because of the source of bias. Rather than coming from the algorithm being given race as an input, or a case of a poorly built or inaccurate model, it was derived from a reasonable modelling choice, which involved the use of cost as a convenient, measurable stand-in for a complex concept, health need, that encoded a real-world inequity into the variable of the model's target. According to the researcher, this is an applicable lesson: when the proxy a model is trained to predict is shaped by unequal treatment or access, a very accurate model can still produce an unfair outcome (Obermeyer et al., 2019).

Subsequent work has extended and complicated this finding at the same time. Morid & Sheng (2025) showed that architectural choices in deep learning cost models can independently affect fairness, which reduces underpayment and overpayment for high-need patients by redesigning the representation of heterogeneous claims data, not just by changing the prediction target. However, according to Zink, Obermeyer, & Pierson (2024), in a colorectal cancer risk-prediction algorithm, which explicitly included race as a model input, there was an improvement in predictive accuracy for Black patients when a key predictor was recorded less reliably for the group. This further complicated any simple rule that clinical algorithms should always exclude race.

➤ *Prediction-Driving Variables*

Across these studies, a recurring finding shows that existing healthcare use is the strongest predictor of future cost, superior to demographic variables like gender or age alone. In the Texas Medicaid study, examining the consistency of a patient's spending from one period to another. Experts show that in healthcare expenditures, patients who were high-cost in a period were largely more likely than the general population to remain high-cost in the subsequent period (Rivera-Hernandez et al., 2022). This is why claims history alone does not suffice for generating a strong prediction (Yang et al., 2018). Corroborating the claims history with clinical detail tends to enhance accuracy, as evidenced by the Japanese national screening study's improvement when claims are combined with lab values and survey data (Osawa et al., 2020).

In practice, this pattern implies that a model designed purely on administrative claims data, without lab or clinical information, can already predict the majority of future high-cost patients because past spending is a strong predictor of future spending (Osawa et al., 2020; Morid & Sheng, 2025). Therefore, a model built only on the total cost already paid for a patient has the tendency to reproduce the patient whom the system has under-served in history.

IV. DISCUSSION

Collectively, this literature makes a contradictory case, where machine learning cost-prediction models are effective and can be destructive, while both can also be true of the high-accuracy model (Olatunji, 2025). Random forest and gradient boosting methods show superior performance to similar

approaches at the task of forecasting the patient who will spend the most on healthcare the following year, with AUCs above 0.87 and 0.91 in another case study. These are strong models by any conventional measure of predictive accuracy.

Meanwhile, the Obermeyer et al. (2019) findings show that accuracy at the specific task is different from the fairness in the outcome meant to be served by the task. In other words, a model can be excellent at predicting cost and wrong about who needs help, especially if need and cost have been allowed to drift apart for some categories of patients because of a healthcare system that has never treated patients equally. This is subtle and more troubling than the inaccuracy of a model because no indicator about the performance metrics of the model would identify the problem. Instead, only a direct audit against an independent measure of actual health revealed it.

This has a direct implication for deploying a cost-prediction model in a U.S. insurance market today. The choice of prediction target goes beyond a technical detail, but remains the most consequential design choice. A model trained to predict illness, clinical risk, or functional decline, using lab values, diagnoses, and validated severity measures, is less likely to inherit cost-driven bias than one trained to predict the amount spent (Dombrowski, Prenner, & Kainz, 2024; Pietrobon et al., 2025). Many of the higher-performing models reviewed in this paper already point in this direction. The best results of the Japanese HNHC study are derived from the combination of clinical data and claims data, not just relying on claims (Osawa et al., 2020). This is consistent with the idea that clinical signal is a more direct and fairer foundation for need-based prediction than billing history.

For health systems and insurers, the practical takeaway is not to neglect cost-prediction modelling, since there is strong evidence on its ability to accurately identify high-burden patients. Instead, it is to audit the models against independent measures of health need, distinguished by race and other demographic groups pre- and post-deployment. Hence, the study by Obermeyer et al. (2019) is widely cited because it showed that such an audit is feasible and able to catch a bias that would be ordinarily missed by accuracy metrics. Standardizing the audit practice is the path toward maintaining the real predictive power of the models and avoiding the failure mode that constitutes the algorithm a destructive tool at scale.

In a recent review of broader algorithmic-fairness literature, experts caution that popular fixes aimed at improving the fairness of healthcare algorithms can sometimes worsen outcomes for patients across ethnic and racial groups, which underscore the need to consistently scrutinize the design choices reviewed in this paper, rather than conduct a one-off correction (Coots et al., 2025).

V. LIMITATIONS

Many of the model-performance studies in the literature, including the highest-accuracy comparison of algorithm types, were conducted outside the U.S., and cost-prediction accuracy can be peculiar to the specific insurance,

billing, and care-delivery structure of a country, limiting direct transfer of exact performance figures to U.S. insurance markets. For example, the audit by Obermeyer et al. (2019) examined a specific commercial algorithm at a health system, while the researchers have debated the underlying mechanism, especially the cost used as a proxy for need, which is common across the wider industry of similar algorithms. Hence, the review cannot confirm the same size of bias in every cost-prediction tool currently being used. Finally, many accuracy-based studies did not report fairness metrics as those extracted by income or race, implying more evidence about the effectiveness of the models to predict cost.

In addition, because much of the newest cited literature, including ICU-focused and pediatric-based studies, was published recently (Nakano et al., 2025; Zhang et al., 2025), longer-term validation and independent replication or transferability of their fairness and accuracy claims are limited.

VI. CONCLUSION

In conclusion, machine learning models have the capability to identify high-cost (high-burden) patients in U.S. insurance markets with documented accuracy, which consistently reached high AUCs within the 0.80 – 0.90 range using claims data only. This is a genuinely useful capability as a small proportion of patients make up the large share of national healthcare spending, and thus, capturing their pattern early helps to create an opportunity for improved and more effective care management. However, the most important lesson derivable from this literature is about the focus of a model, particularly the task asked to predict. When cost is substituted for need, and from history, cost has been shaped by unequal access to care, an accurate cost model can become unfair and inaccurate without showing a sign in its performance statistics. Therefore, the way forward is not to lose trust in the tools, but to maintain that every cost-prediction model used in a U.S. insurance market is audited against independent measures of actual health need, extracted by income and race.

REFERENCE

- [1] Bruyneel, A., Larcin, L., Martins, D., van den Bulcke, J., Leclercq, P., & Pirson, M. (2023). Cost comparisons and factors related to cost per stay in intensive care units in Belgium. *BMC Health Services Research*, 23, 986.
- [2] Chang, C. F., Mirvis, D. M., & Mahmood, A. (2026). The high cost of American health care: Understanding the deeper roots of the crisis. *Health Care Analysis*, 34(1), 1–12.
- [3] Coots, M., Linn, K. A., Goel, S., Navathe, A. S., & Parikh, R. B. (2025). Racial bias in clinical and population health algorithms: A critical review of current debates. *Annual Review of Public Health*, 46, 507–523.
- [4] Crowley, R., Daniel, H., Cooney, T. G., Engel, L. S., & Health and Public Policy Committee of the American College of Physicians. (2020). *Envisioning a better US health care system for all: Coverage and cost of care*. *Annals of Internal Medicine*, 172(2, Suppl.), S7–S32.
- [5] Delanerolle, G., Bouchareb, Y., Katsikopoulos, K. V., & Phiri, P. (2026). *Data science in healthcare: A complete guide*. CRC Press.
- [6] Dickman, S. L., Gaffney, A., McGregor, A., Himmelstein, D. U., McCormick, D., Bor, D. H., & Woolhandler, S. (2022). Trends in health care use among Black and White persons in the US, 1963–2019. *JAMA Network Open*, 5(6), e2217383.
- [7] Dombrowski, M., Prenner, A., & Kainz, B. (2024). Bias assessment and data drift detection in medical image analysis: A survey (arXiv:2409.17800). *arXiv*. <https://arxiv.org/abs/2409.17800>
- [8] Hajat, C., Siegal, Y., & Adler-Waxman, A. (2021). Clustering and healthcare costs with multiple chronic conditions in a US study. *Frontiers in Public Health*, 8, 607528.
- [9] Hutchins, E. J. (2025). *Evaluating the additive value of machine learning in clinical risk prediction: A comparative study across three patient cohorts* [Doctoral dissertation, University of California, Los Angeles].
- [10] Jones, C. H., & Dolsten, M. (2024). Healthcare on the brink: Navigating the challenges of an aging society in the United States. *npj Aging*, 10(1), 22.
- [11] Langenberger, B., Schulte, T., & Groene, O. (2023). The application of machine learning to predict high-cost patients: A performance comparison of different models using healthcare claims data. *PLOS ONE*, 18(1), e0279540.
- [12] Maisog, J. M., Li, W., Xu, Y., Hurley, B., Shah, H., Lemberg, R., ... & Gutfraind, A. (2019). Using massive health insurance claims data to predict very high-cost claimants: a machine learning approach. *arXiv preprint arXiv:1912.13032*.
- [13] Mo, C. (2026). *Optimizing care management targeting: A machine learning comparison of regression, classification, and ranking approaches for identifying future high-cost patients*.
- [14] Morid, M. A., & Sheng, O. R. L. (2025). Healthcare cost prediction for heterogeneous patient profiles using deep learning models with administrative claims data. *Information Systems Research*, 36(4), 1968–1992.
- [15] Mwakinyali, F., Kateule, R., & Tunga, M. (2024). Forecasting pediatric medical expenses using machine learning: A case study of the Toto Afya Card in Tanzania. *Journal of ICT Systems*, 2(2), 55–66.
- [16] Nakano, F. K., Van Aerde, N., Coppens, G., Vanhorebeek, I., Vens, C., Van den Berghe, G., & Güiza Grandas, F. (2025). Development and validation of a machine learning model for early prediction of intensive care unit acquired weakness. *Intensive Care Medicine Experimental*, 13, 98.
- [17] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

- [18] Olatunji, A. O. (2025). Machine learning in business process optimization: A framework for efficiency and decision-making. *Journal of Basic and Applied Research International*, 31(2), 18–28.
- [19] Osawa, I., Goto, T., Yamamoto, Y., & Tsugawa, Y. (2020). Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data. *npj Digital Medicine*, 3, 148.
- [20] Pietrobon, R., Machiavelli, A., Paulsen Rodrigues, L., Agrey, A., Nkeangnyi, L., Zechia, G., ... Teixeira, L. (2025). The bias assessment method: Evaluating racial inequities in clinical risk prediction models (motivated by Obermeyer et al.'s analysis of cost-driven labeling in healthcare algorithms). SSRN. <https://ssrn.com/abstract=5861522>
- [21] Rahman, Y., & Dua, P. (2025). A machine learning framework for predicting healthcare utilization and risk factors. *Healthcare Analytics*, 8, 100411.
- [22] Rivera-Hernandez, M., Kumar, A., Chou, L. N., Keeney, T., Ferdows, N., Karmarkar, A., & Ottenbacher, K. (2022). Healthcare utilization and costs among high-need and frail Mexican American Medicare beneficiaries. *PLOS ONE*, 17(1), e0262079.
- [23] Shockney, L. D. (2021). Chronic disease and complex care navigators: A scoping review.
- [24] Zhang, P., Zhu, B., Chen, X., & Wang, L. (2025). Predicting high-need high-cost pediatric hospitalized patients in China based on machine learning methods. *Scientific Reports*, 15, 16006.
- [25] Zink, A., Obermeyer, Z., & Pierson, E. (2024). Race adjustments in clinical algorithms can help correct for racial disparities in data quality. *Proceedings of the National Academy of Sciences*, 121(34), e2402267121.