QUBAHAN ACADEMIC JOURNAL
VOL. 5, NO. 4, December 2025
https://doi.org/10.48161/qaj.v5n4a2198

Data-Driven Risk Stratification for High-Cost Care Management: An Empirical Evaluation of Generalized and Regularized Models

Eslam Abdelhakim Seyam [1*]

[1] Department of Insurance and Risk Management, College of Business, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11564, Saudi Arabia.

* **Corresponding author:** isiam@imamu.edu.sa.

ABSTRACT: This paper examines data-driven methods for the identification of high-cost patients in European health systems by assessing predictive accuracy and interpretability in generalized and regularized statistical models. We learn binary classification problems from a big health insurance dataset to identify individuals in the upper 2.5% and upper 5% of overall healthcare spending. Three modeling paradigms-Generalized Linear Models (GLM), Generalized Additive Models (GAM), and LASSO regression-are used and contrasted in terms of predictive accuracy as well as practical interpretability. We find that GAM consistently outperforms, yielding highest F1 values and lowest log loss, in capturing nonlinear associations in health care consumption better than GLM or LASSO. Frequency of surgeries, hospitalizations, and duration of insurance coverage prove to be key determinants of high-cost status, while demographic attributes like gender exert a moderate impact. The comparisons highlight the potential of utilizing interpretable yet adaptable models to enable proactive, risk-based interventions. By presenting evidence of predictive accuracy vs. interpretability trade-offs, the paper aids more efficient high-cost care management, providing pragmatic advice to European health systems to efficiently allocate assets in light of avoidable health care spending.

Keywords: high-cost patients; European health systems; health insurance data; predictive modeling; generalized linear models (GLM); generalized additive models (GAM); lasso regression.

I. INTRODUCTION

Healthcare expenditure is notoriously unevenly distributed across populations, with a small fraction of patients accounting for a disproportionate share of costs. In the United States, for example, the top 5% of spenders are responsible for nearly half of all health care spending [1]. Similar skewed cost distributions are observed in high-income countries worldwide, including Europe, where health systems face the challenge of managing "high-cost, high-need" patients within constrained budgets. Throughout this paper, we use "high-cost patients" to refer specifically to individuals whose total healthcare expenditures place them in the uppermost percentiles of the cost distribution (top 2.5% or 5% in our analysis). When discussing the broader literature, we use "high-need, high-cost (HNHC)" to reflect the common designation for patients with both elevated costs and complex care needs. The concentration of spending in a very small group of patients has been well documented in European contexts such as the United Kingdom and the Netherlands, underscoring that this phenomenon is not unique to the U.S. [2]. This heavy-tailed distribution of health costs motivates a critical focus on those high-cost patients as a strategy for improving health system efficiency and sustainability. High-cost patients often suffer from multiple chronic conditions and complex health needs, which drive their frequent use of expensive services. Research indicates that these individuals are more likely to experience fragmented or suboptimal care quality and safety issues, suggesting ample room for improvement in how their care is managed [2].

A profile of "high-need, high-cost" patients show they commonly have multiple comorbidities, functional limitations, or other frailties that complicate their care [3]. Many are elderly patients with conditions such as diabetes, heart failure, mental health disorders, or combinations thereof, which require continuous management. Because their care needs are complex and ongoing, unmanaged high-cost patients can cycle repeatedly through emergency departments and hospital admissions, incurring substantial costs. Targeting this small cohort for improved care management, therefore, holds significant promise: by enhancing care coordination, preventive interventions, and chronic disease management for these patients, health systems may achieve better health outcomes while containing unnecessary expenditures. Indeed, early efforts in health policy have posited that focusing on the sickest few patients could yield outsized gains in both cost containment and quality of care [1]. European health systems, many of which provide universal coverage and operate under fixed budgets, have a particularly strong incentive to identify and proactively manage high-cost patients. In England, for instance, roughly 30% of the population lives with at least one long-term condition, and care for these patients accounts for about 70% of health and social care spending [4]. Faced with ageing demographics and a rising burden of chronic disease, numerous European countries have developed national strategies for high-cost care management. These strategies often draw on the paradigm of risk stratification and early intervention: stratifying the population by risk allows health services to pinpoint the small segment of patients at greatest risk of becoming (or remaining) very high-cost users.

By intervening early-through enhanced primary care support, multidisciplinary case management, or integrated care pathways-health systems aim to prevent complications and avoidable hospitalizations in this group. Such approaches align with the widely adopted "pyramid" model of care (exemplified by Kaiser Permanente) that has influenced European chronic care management, where the top tier of the pyramid (a few high-risk patients) receives intensive, coordinated care. Policymakers and researchers across Europe have increasingly turned to data-driven tools to enable this targeted approach. For example, risk stratification algorithms and predictive modeling are now employed in countries like Spain, the UK and others to forecast which patients are likely to incur extreme costs, so that preemptive care plans can be put in place [5]. This rise of data driven predictive analytics in European health systems reflects a broader trend: leveraging electronic health records, claims data, and advanced statistical models to improve resource allocation and care delivery for those who need it most.

Against this backdrop, effective risk stratification becomes a linchpin for high-cost care management. There is a growing body of evidence and practical experience in Europe suggesting that well-calibrated predictive models can identify future high-cost individuals with reasonable accuracy, informing more efficient care management programs [6]. Early initiatives, such as the Combined Predictive Model in England and similar programs in Spain's Basque Country, have demonstrated both the feasibility and potential benefits of integrating predictive risk tools into routine care [5]. These tools enable health providers to move from a reactive stance-treating complications after they arise-to a proactive strategy of managing health risks before they escalate into acute (and expensive) events. In doing so, health systems not only aim to contain costs but also to improve patient outcomes and experiences, particularly for vulnerable groups with complex care needs. The emphasis on high-cost patient programs thus aligns with broader goals of healthcare efficiency and equity, ensuring that resources are directed to the patients who generate (and arguably need) the most care.

In this paper, we contribute to the literature on data-driven risk stratification for high-cost care management by providing an empirical evaluation of generalized and regularized predictive modeling approaches in a European healthcare context. Importantly, our analysis quantifies just how extreme the cost concentration can be: we find that a mere 2.5% of patients account for approximately 36.56% of total healthcare expenditures, while the top 5% of patients account for fully 52.95% of costs. These original findings, based on our comprehensive patient-level data set, reaffirm the critical importance of focusing on the high-cost cohort. They also underscore the value of robust predictive risk stratification the better to flag these patients prospectively as a means to guide efficient intervention and resource allocation. In the following sections, we situate these findings in the context of existing research and discuss how advanced predictive models can enhance high-cost care management. Our results shed new light on the potential gains from employing modern analytical techniques in European health systems to identify and manage high-cost patients, ultimately aiming to improve system-wide cost-effectiveness and patient care outcomes.

Our analysis demonstrates that Generalized Additive Models (GAM) consistently outperform both standard Generalized Linear Models (GLM) and LASSO regression across multiple performance metrics, primarily due to GAM's ability to capture nonlinear relationships between utilization patterns and cost outcomes while maintaining interpretability through visual inspection of smooth functions. The remainder of this paper is structured as follows: Section 2 reviews existing literature; Section 3 presents methodology; Section 4 presents empirical findings; and Section 5 discusses implications and future research.

II. LITERATURE REVIEW

Health expenditures in many health systems are highly skewed, with a small fraction of patients accounting for a disproportionate share of costs. For example, the top 5% of patients can incur roughly 50% of total healthcare expenditures [7]. High-need, high-cost patients have consequently been identified as an urgent priority for health policy interventions. Recent European evidence confirms this pattern: Grout et al. (2024) report that up to 80% of EU healthcare costs are attributable to chronic disease patients [8], while concentration rates vary across European countries from 15% to 33% for the top 1% of spenders [2]. In response, healthcare researchers have focused on predictive risk stratification models to identify future high-cost individuals so that care management resources can be targeted effectively. Early studies in this domain typically employed traditional statistical models-especially generalized linear models (GLMs) and related regression approaches - to predict healthcare costs or classify patients above a cost threshold [9]. For instance, logistic regression models using demographics, diagnoses, and prior utilization were shown to moderately predict the risk of becoming a high-cost user in settings like Ontario's health system [10, 11]. However, such linear models have limitations in capturing complex nonlinear relationships and interactions in the data. This recognition has led to growing interest in machine learning techniques that promise higher predictive accuracy, albeit often at the expense of interpretability.

A variety of predictive modeling strategies have been considered for high-cost patient identification, from traditional statistical models to more advanced machine learning techniques. These methods have been directly compared in recent empirical studies on large real-world claim datasets, providing information about their relative performance. Systematic review and comparison by [12] used multiple methods to forecast next year's cost for a health insurer's claims data (about $\sim 90{,}000$ individuals) and compared them. Tree ensemble-based methods, especially gradient boosting, reported optimum overall predictive accuracy (measured in terms of metrics such as R^2 and C-statistic) for ordinary populations. Surprisingly, however, in the specific case of finding highest-cost individuals (e.g. top cost decile), less sophisticated models performed very competitively: a regularized linear regression (Ridge) and a multilayer neural network topped that extreme subgroup. It indicates that though high-end non-linear methods may lead in broad accuracy, more basic generalized models with regularization can perform strongly in singling out cost outliers, perhaps because they can generalize without overfitting over rare high-cost observations [12].

Overall, the literature shows that machine learning methods (random forests, boosting, neural nets) consistently outperform traditional GLMs in predictive accuracy for healthcare cost stratification, often by a meaningful margin, especially when rich claims data are available. For example, a recent study in Germany applied logistic regression, random forest (RF), gradient boosting, and a deep neural network to statutory insurance claims for over 20,000 individuals [13]. The tree-based models achieved the highest discrimination in identifying next-year top 5% cost patients, with RF reaching an AUC of about 0.88 on test data, compared to ~ 0.84 for the neural network and logistic regression. All approaches had acceptable performance (AUC ≥ 0.8), but the ensemble methods were statistically significantly better, highlighting the value of non-linear interactions and ensembles in capturing the complex drivers of cost. A variety of predictive modeling strategies have been compared in recent empirical studies. Langenberger et al. (2023) applied random forest (RF), gradient boosting machine (GBM), artificial neural network (ANN), and logistic regression to German statutory insurance claims for over 20,000 individuals [14]. The tree-based models achieved highest discrimination in identifying next-year top 5% cost patients, with RF reaching an AUC of about 0.88, compared to approximately 0.84 for neural network and logistic regression. Similarly, a study using French national health insurance data (over 500,000 individuals) found that a random forest model explained the most variance in individual costs ($R^2 \approx 0.48$), substantially outperforming a conventional GLM ($R^2 \approx 0.35$) on the same features [15]. The neural network in that study performed intermediate ($R^2 \approx 0.32$), reinforcing that tree ensembles often have an edge in tabular claims data settings.

While it is critical to aim for high predictive accuracy, there is an equally critical requirement to balance performance with interpretability in clinical implementation. Stakeholders like clinicians and care managers need understandable reasoning about why a given patient is high-risk for cost, to enable trust-building as well as to determine intervention strategies. Models that are interpretable, such as GLMs or decision trees, have explicit relationships among predictors and outcome, but complex models (decision trees, boosted trees, deep neural nets) tend to be "black boxes." Explainable AI (XAI) methods like SHAP (Shapley Additive exPlanations) have been used to gain insights from blackbox models. For instance, Orji and Ukwandu (2024) used SHAP values to rank high-cost drivers in a boosted tree ensemble [16]. However, even with XAI tools, obtaining a comfortable trade-off remains difficult [17, 18]. The literature documents this trade-off. For example, Vimont et al. (2022) determined that while a random forest provided the highest cost prediction in French claims, a GLM model was "well suited" if what you want to know is what individual predictors contribute. That is, if interpretability is most important (e.g. to understand which diagnoses, drugs or utilization factors drive risk), lower accuracy (albeit lower) generalized models such as these might be preferable. But if what you want, solely, to identify accurately as many correct patients as possible (e.g. for outreach enrollment in care management), more complex models have an advantage [14].

One of the areas of research that continues to expand is therefore the use of eXplainable AI (XAI) methods to gain insights from black-box models. For instance, Orji and Ukwandu (2024) used SHAP (Shapley Additive Explanations) values and Individual Conditional Expectation plots to a boosted tree ensemble used to predict insurance cost, in order to rank high-cost drivers and visualize how variations in a feature (such as a chronic condition flag or age) impact predicted cost for a given individual [16]. Similarly, Langenberger et al. (2023) used a SHAP analysis of their random forest model to determine which most influential features drive high-cost prediction in a German dataset. These methods increase transparency in machine learning models, in turn closing the gap between accuracy and interpretability. However, obtaining a comfortable trade-off is difficult. Simpler models (e.g. LASSO-regularized regression) produce easily interpretable risk scores and have been utilized to choose a limited number of predictors (e.g. a few chronic conditions or prior cost flags) for risk stratification models [19]. Such models can sometimes achieve similar performance to more complex models but with much more transparency. More broadly, existing evidence indicates no single model is always to be preferred - there's a trade-off between sheer predictive power (where ensemble and network models lead) and understandability/explainability (where generalized linear and additive models lead). Many authors recommend combining advanced models with interpretability tools, or with hybrid strategies (e.g. rule-based segmentation with subsequent machine learning in each segment) to balance both needs.

A further key consideration is predictive model generalizability to other health systems and populations. The majority of studies to date have been carried out in North America or European contexts in a given insurance claims base, and findings might not apply directly in other contexts. Cross-national comparisons identify that high-cost patient profile as well as health cost structure can differ by system widely [2]. For instance, percentage of costs attributable to top 1% ranged from 15% in one to a high of 33% in another. This type of variation would mean that a model built from a single nation's data, in turn, might not perform as similarly in any other place in the absence of adaptation. Even in Europe, by way of illustration, a (German, say) claims-derived or (UK, say) claims-derived risk stratification model might require recalibration in an equivalent German or UK setting because coding, practice style, as well as health profiles of populations differ. This points to a requirement for local validation studies. Despite great developments in predictive modeling for high-cost care management, notable gaps exist. First, most studies have been carried out under the limitations of single country datasets, raising concerns over extensibility to different European populations. Second, there exists an ongoing trade-off between predictive accuracy and interpretability, with most precise machine learning algorithms compromising transparency, which can create hindrances in clinical practice where decision support needs to be understandable and credible.

The current research aims to bridge these gaps by empirically comparing a set of generalized and regularized models (GLM, GAM, LASSO) based on real-world claims information from a European health system. By analyzing a single European dataset, this research conducts a rigorous internal comparison of modeling strategies while emphasizing the importance of careful consideration when transporting models between regions. Our comparison focuses not only on predictive accuracy for classifying high-cost patients but also on interpretability and transportability of each model. This research helps to elucidate which modeling strategies offer optimal performance versus explainability in European health insurance data, informing more contextually appropriate and actionable risk stratification strategies.

III. METHODOLOGY

1. DATA DESCRIPTION AND PREPROCESSING

The dataset contains anonymized health insurance records from a private European insurer, covering 176,032 individuals between 2012-2018. Policyholders were 51.3% female with a mean age of 37.5 years, and average policy duration of 2.76 years. Thirteen variables capturing demographics, policy characteristics, and healthcare utilization across major service categories were included. After preprocessing, 123,029 observations were allocated to the training set, 26,502 to validation, and 26,501 to testing following a stratified 70-15-15 split. Fully anonymized data were handled in compliance with GDPR. As the study used de-identified secondary administrative data, formal ethical approval was not required under institutional guidelines. Approximately 3.2% of service-use values were MCAR (Little's test $p = 0.31$) and were imputed as zero, reflecting likely non-use. Sensitivity checks using alternative imputations produced minimal changes ($< 2\%$).

A total cost variable was constructed by aggregating all cost components, and highcost cases were defined using the top 2.5% and 5% thresholds. Service-use measures were normalized by policy duration, and categorical variables were encoded numerically. The resulting dataset supported the development and evaluation of predictive models. Cost distribution analysis confirmed substantial right skewness, with a small fraction of individuals accounting for a large share of expenditures. This reinforces the importance of identifying high-cost users for targeted resource allocation. Several limitations apply: reliance on a single private insurer limits generalizability; uncovered out-of-pocket spending is absent; and inherent selection effects in privately insured populations may introduce bias.

Table 1. Variable definitions.

Variable	Definition	Statistical Type
high_cost	Indicator variable equal to 1 if total cost exceeds a defined high-cost threshold, 0 otherwise.	Binary
age_at_inception	Age of the policyholder at the time of insurance policy inception.	Continuous (Ratio)
gender	Encoded as 1 for female, 0 for male.	Binary
relation	Encoded categorical variable indicating relationship to the policyholder (such as owner, partner, child).	Categorical
policy_years	Duration of the insurance policy coverage in years.	Continuous
avg_num_analysis	Average annual number of laboratory analysis procedures per policyholder.	Continuous
avg_num_dentistry	Average annual number of dental procedures.	Continuous
avg_num_diagnostics	Average annual number of diagnostic procedures (such as imaging).	Continuous
avg_num_endoscopy	Average annual number of endoscopy procedures.	Continuous
avg_num_hospitalizat	Anserage annual number of inpatient hospitalizations.	Continuous
avg_num_mammography	Average annual number of mammography procedures.	Continuous
avg_num_operations	Average annual number of surgical operations.	Continuous

avg_num_visits	Average number of outpatient or clinic visits annually.	Continuous

2. GENERALIZED LINEAR MODEL (GLM)

The generalized linear model (GLM) is a versatile extension of standard linear regression in which target Y can have an exponential family distribution (e.g. Poisson, Bernoulli) instead of Gaussian [20]. The three elements in a GLM are: (i) a random component defining Y_i 's distribution (mean μ_i and variance $\operatorname{Var}(Y_i)$ as a function of μ_i), (ii) a systematic component (linear predictor) $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ from a set of features $\mathbf{x}_i$, (iii) a monotonic, smooth link function $g(\cdot)$ relating these through $g(\mu_i) = \eta_i$ [21].

For instance, in logistic regression (a binary outcomes GLM), $g(p) = \log p/(1-p)$ as a logit link guarantee that predicted probability $p = \mu_i$ remains in [0,1]. The canonical log-likelihood of a GLM under independent observations $\{y_i\}_{i=1}^n$ is:

$$\ell(\beta) = \sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{\phi} + \text{ constant }, \tag{1}$$

where θ_i is the canonical parameter related to μ_i (for canonical links, $\theta_i = \eta_i$) and ϕ is a dispersion parameter [20]. In the special case of a Bernoulli ($0/1$) outcome with $\pi\_i = P(Y\_i = 1 \mid \mathbf{x}\_i), \theta\_i = \log[\pi\_i/(1-\pi\_i)[$, and Equation (1) reduces to the usual logistic regression log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)], \text{ with } \pi_i = g^{-1}(\mathrm{x}_i^T \beta) \tag{2}$$

The Generalized Linear Model (GLM) framework, originally formalized by Nelder and Wedderburn [20], brought together many traditional models (linear regression, logistic/probit, Poisson regression, etc.) by employing iterative maximization of likelihood (usually through iteratively reweighted least squares) to estimate β [22]. Key assumptions are that each Y_i is conditionally independent given x_i and follows a known exponential-family distribution with correctly specified mean function $E(Y_i \mid \mathrm{x}_i) = \mu_i$. The link $g(\mu)$ is typically chosen as the canonical link (e.g. logit for binary, log for Poisson) to simplify interpretation and maximize efficiency.

Interpretability and Relevance: GLMs can be interpreted relatively easily since the linear predictor $\eta_i = \mathrm{x}_i^T \beta$ indicates that each variable x_j has a multiplicative impact upon the mean on the scale of the link. For example, in logistic regression, the coefficient β_j represents the log-odds ratio per unit increase in x_j for outcome Y controlling for other variables. This interpretability along with ease of the linear expression make GLMs a common baseline for binary classification problems, including health insurance analytics. When predicting high-cost claimants, a GLM (logistic regression) can explicitly estimate the probabilities of an individual being high-cost, and its coefficients emphasize which factors of risk (age, diagnoses, etc.) increase or decrease those probabilities [23].

Though, by their assumption of a linear relationship (on the link scale), they cannot naturally learn complex nonlinear interactions or effects without feature engineering, they are computationally cheap and less likely to overfit with approximately linear true relationships. Indeed, despite more sophisticated machine learning methods increasingly becoming available, GLMs continue to find widespread usage as an interpretable baseline; for instance, recent research in health cost prediction continues to use logistic regression as a baseline, highlighting its continued validity even as more advanced machine learning techniques become available [24].

3. GENERALIZED ADDITIVE MODEL (GAM)

The generalized additive model (GAM) expands upon the GLM by fitting non-linear functions between each feature and the outcome but retaining additivity for ease of interpretation [25]. The linear prediction η_i in a GAM is replaced by a sum of functions of the covariates. Mathematically, for m features, a GAM assumes:

$$g(\mu_i) = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_m(x_{im}) \tag{3}$$

where α is an intercept, and each $f_j(x_{ij})$ represents a non-parametric smooth function that describes the possible nonlinear impact of predictor j on the outcome. The link function $g(\cdot)$ and Y_i distribution is selected as in a GLM, so GAMs can accommodate binary outcomes (logistic GAM), counts (Poisson GAM), etc., with a likelihood-based approach as in GLMs. The f_j functions are often represented through basis expansions (such as splines or radial basis functions) and estimated by maximizing a penalized log-likelihood. Specifically, if $\ell(\alpha, f_1, \dots, f_m)$ is a log-likelihood of the data as in Equation (1) or (2) for a binary GAM), then the choice of fitting criterion includes adding a roughness penalty to each smooth component:

$$Q(\alpha, f_1, \dots, f_m) = \ell(\alpha, f_1, \dots, f_m) - \sum_{j=1}^{m} \lambda_j J(f_j) \tag{4}$$

where $J(f_j)$ measures the wiggliness (complexity) of f_j (e.g. $J(f_j) = \int \left[f_j''(x)\right]^2 dx$ for a spline) and $\lambda_j \geq 0$ are smoothing parameters that control the trade-off between fit and smoothness [26]. In practice, the λ_j 's are selected by methods such as cross-validation, restricted maximum likelihood, or generalized cross-validation in order to make each $f_j(x)$ as smooth as possible with minimal loss of substantial model fit. First implementations estimated GAMs by a backfitting algorithm iteratively updating each f_j with the rest held constant. Modern methods (such as by means of the mgcv R package) directly solve the penalized scoring equations, which provides efficient joint estimation of all f_j 's as well as their levels of smoothness.

Assumptions and Statistical Properties: GAMs suppose that the true mean structure can be represented by an additive sum of smooth effects. This implies GAMs ignore any higher-order interaction terms among predictors unless those explicitly enter (e.g. through tensor product smooths or interaction terms). Every smooth function f_j is generally assumed to have a high enough level of smoothness (e.g. to be twice differentiable) in order for the concept of wiggliness to make proper sense. Besides the additivity assumption, GAMs inherit the remaining GLM assumptions: independent observations, as well as a correctly specified distribution/link for Y. When these hold, the GAM can be interpreted as a large GLM with basis function covariates, with inference for the smooth terms (confidence bands, testing) being done by approximate degrees of freedom for each f_j [27].

Interpretability and Complexity: One of the key benefits of GAMs is their interpretability: even though each $f_j(\cdot)$ can be a complicated non-linear curve, it can be plotted to learn about how feature x_j affects the outcome on the log-odds scale. For a logistic GAM for claim risk, for instance, it could be plotted to visualize how log-odds of high-cost claimant changes as a smooth function of age. This is more informative than a black-box pure model, yet more flexible than a linear model. Precisely, GAMs have now come to be seen as a top choice among interpretable machine learning models, which have predictive accuracy as well as transparency about feature effects [17].

The complexity of a GAM is greater than that of a standard GLM since each f_j can make use of many degrees of freedom (basis functions) to represent non-linearity. The added smoothing penalties λ_j, however, serve to effectively regularize such complexity, shrinking f_j toward a linear (or even constant) curve in instances where data fail to strongly support a more complex specification. This regularization parallels shrinkage in penalized regressions and is fundamental to make GAMs useful in practice. In insurance, health economics, and other areas, GAMs have been found useful for fitting outcomes that have nonlinear relationships with predictors (e.g. health care expenditure as a function of age often have a nonlinear increase for higher ages). With GAMs, analysts have been able to make more accurate predictions than with linear models while still meeting the requirement for explainability (e.g. illustrating regulators' impact of each rating factor upon insurance premiums in a smooth curve, as opposed to a constant coefficient) [28]. The pioneering efforts of Hastie and Tibshirani [25] in the 1980s introduced GAMs, since which time research over the past ten years has continued to develop and extend this methodology and use it in large-scale applications (such as current variations such as gradient boosting GAMs and shape-constrained GAMs) [29].

4. LASSO REGRESSION

The LASSO (Least Absolute Shrinkage and Selection Operator) regularization technique adds to a regression model an ℓ_1 penalty to implement automatic variable selection along with coefficient shrinkage [30]. For a generalized linear model (such as logistic regression for binary classification), the LASSO estimator $\hat{\beta}$ can be defined as the solution to the constrained optimization.

$$\hat{\beta} = \arg\min_{\beta} \{-\ell(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|\}, \tag{5}$$

where $\ell(\beta)$ is the log-likelihood of the unpenalized model such as Equation (2) for logistic regression) and $\lambda \geq 0$ is a tuning parameter controlling the strength of the L^1 penalty. Equivalently, LASSO can be viewed as maximizing the likelihood subject to a budget constraint $\sum_j |\beta_j| \leq t$ for some t (with a one-to-one correspondence between t and λ) [?]. The ℓ_1 penalty shrinks coefficient estimates towards zero, and in turn, with sufficiently large λ, some coefficients will equal zero. This distinguishes LASSO from ridge regression (which employs an ℓ_2 penalty and only shrinks coefficients, never setting any to zero). Bayesianly, the LASSO can be viewed as a maximum a posteriori estimate with a doubleexponential (Laplace) prior over coefficients β_j with heavy tails and a spike at zero to promote sparsity [31].

Assumptions and Tuning: The LASSO, as other penalized regression, assumes that the true underlying model is sparse or can be approximated by a sparse subset of predictors. For high-dimensional settings (i.e., where p is large, often $p \gg n$), LASSO builds upon the assumption that most of the predictors have a zero or negligible impact, and it attempts to pick out the salient ones. There exist technical conditions (such as the restricted eigenvalue or coherence conditions on the design) under which LASSO consistently selects correct variables as n increases [32].

In practice, λ is key and is often chosen by cross-validation or information criteria: a higher λ provides a more parsimonious fit (more zeros), while a lower λ provides a fit closer to the complete maximum likelihood estimates. The λ which is chosen often achieves a good bias-variance trade-off, minimizing out-of-sample prediction loss [?]. For a binary classification task such as high-cost claim prediction, it's possible to use cross-validated deviance or AUC to choose λ which achieves a balance between model complexity and prediction accuracy. Another assumption behind LASSO is that all predictors have been measured from similar scales or have been standardized; otherwise, penalty $\sum_j |\beta_j|$ will lean towards choosing variables with smaller scales. Standard practice, therefore, is to standardize continuous features prior to applying LASSO.

Interpretability and Use in Classification: The LASSO is sometimes celebrated to add interpretability to intricate modeling applications [33]. By setting to zeros less informative features, it generates a more interpretable and understandable model. For healthcare cost classification, a logistic LASSO model could begin with possibly hundreds of potential predictors (such as diagnostic codes, demographic factors) and reduce to a much smaller subset which significantly informs prediction of high-cost cases. This sparsity allows for effectively explaining key drivers of prediction to stakeholders (such as clinicians or underwriters), resolving a frequent criticism of over-parameterized models.

Furthermore, by shrinking the number of predictors, LASSO can combat overfitting and often enhances the generalization performance, in particular if p is large compared to n [34]. It should however be noted that LASSO's shrinkage aggressively induces a representation bias in estimated coefficients-while selected variables tend to be relevant, their coefficients shrink toward zero. This can subtly impede pure effect size interpretability (e.g. a LASSO coefficient's size isn't an unbiased estimator of the log-odds ratio in reality), though refitting an unpenalized model to the selected features or making use of adaptive LASSO implementations can reduce this. In the last ten years, the LASSO has emerged as a standard in high-dimensional statistical learning [35]. Both successful applications as well as an abundant supporting theory have contributed to its popularity: numerous studies have examined its consistency, sampling asymptotic, and variants (group LASSO, fused LASSO) for different forms of data.

From a practical perspective, efficient algorithms for coordinate descent made it convenient to implement LASSO even for large datasets, and success in competitions, applications (such as genomics selection, text classification, etc.), as well as its use in established centers of high-dimensional statistical learning (biostatistics, actuarial science), established its credentials. For instance, a current clinical research project might employ LASSO logistic regression to winnow a multitude of patient features down to a few salient contributors to high medical expense, resulting in a sparse predictive model that clinicians can inspect, validate, and interpret. Methodologically, research continues to innovate in terms of LASSO's usage recent reviews address, inter alia, choosing tuning parameters optimally as well as uncertainty in LASSO models [19]. To sum up, the LASSO offers a principled mechanism to regularize as well as select features in the standard GLM framework, making it a prime candidate for binary classification problems with a multiplicity of potential features (such as prediction of high-cost claims). It was introduced by Tibshirani in 1996, a turning point in statistical learning, and continues to represent a leading modern regression method, closing the gap, as it does, between interpretable simple models and high-dimensional, adaptable learning modes.

IV. RESULTS AND DISCUSSION

1. DESCRIPTIVE STATISTICS

The descriptive statistics in Table 2 suggest great variation in health care utilization among insured individuals. The mean age at policy start was about 37.46 years (SD = 12.10), reflecting a mature age profile with a considerable span from infancy to old age. The policy duration averaged out at 2.76 years (SD = 1.84), though a high percentage of policies had a relatively short duration, as can also be seen from median as well as first quartile values equalling 1.5 years.

Most medical-service variables, including mean annual procedures and visits, have strongly right-skewed distributions with means many multiples higher than medians and third quartiles equal to zero. The mean number of analysis procedures, for example, was 0.067, while median and Q3 both equaled zero, suggesting that most persons received none of these in a given year, but a minority showed very high usage (such as many as 11.98 for analyses and 22.62 for hospitalizations). This skewness provides rationale for the use of robust modeling methods that can handle non-normal as well as zero-inflated distributions, particularly in detecting high-cost subgroups. These distributional features justify the use of nonlinear and regularized regression methods (such as GAM and LASSO) used subsequently in the analysis to account for heterogeneity in health care consumption.

Table 2. Summary statistics of continuous variables ($n = 176{,}032$).

Variable	Min	Q1	Median	Mean	SD	Q3	Max
Age at Inception (Years)	0.00	31.18	38.12	37.46	12.10	45.88	73.18
Policy Duration (Years)	0.01	1.50	1.50	2.76	1.84	5.00	5.75
Avg. Analysis Procedures	0.000	0.000	0.000	0.0669	0.289	0.000	11.98
Avg. Dentistry Procedures	0.000	0.000	0.000	0.0538	0.189	0.000	6.65
Avg. Diagnostics Procedures	0.000	0.000	0.000	0.0210	0.118	0.000	3.99
Avg. Endoscopy Procedures	0.000	0.000	0.000	0.00287	0.0388	0.000	2.00
Avg. Hospitalizations	0.000	0.000	0.000	0.00939	0.167	0.000	22.62
Avg. Mammography Procedures	0.000	0.000	0.000	0.0302	0.145	0.000	3.69
Avg. Operations	0.000	0.000	0.000	0.00054	0.0252	0.000	5.56
Avg. Visits	0.000	0.000	0.000	0.102	0.339	0.000	7.32

Min refers to the minimum observed value, Q1 indicates the first quartile, Median represents the middle value of the distribution, Mean denotes the arithmetic average, SD indicates the standard deviation, Q3 refers to the third quartile, and Max represents the maximum observed value in the sample.

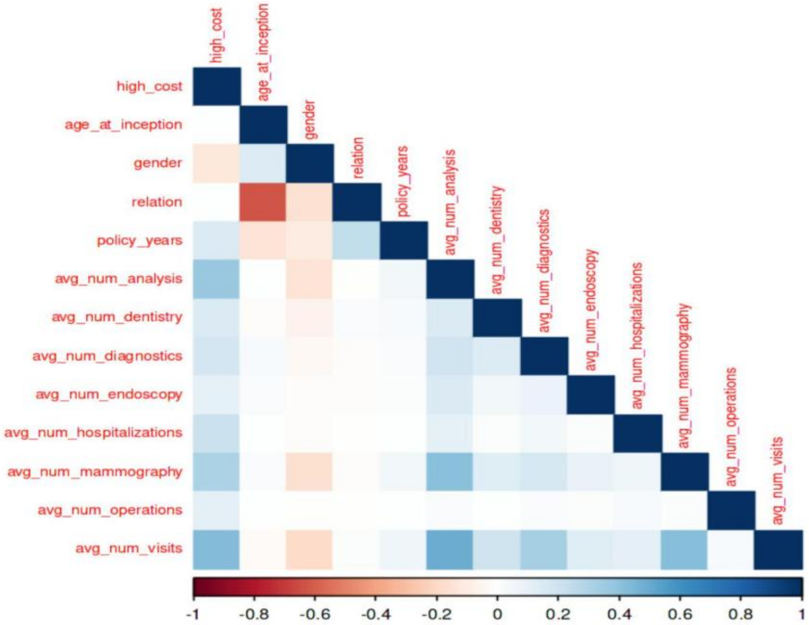

FIGURE 1. Pearson correlation matrix for key continuous and binary variables. Stronger correlations are shown in darker shades. Negative correlations are in red and positive in blue.

Figure 1 illustrates the Pearson correlation matrix of primary continuous and binary variables employed in predictive modeling. The correlations among predictors overall are low to moderate, signifying little multicollinearity, which improves model interpretability as well as stability. As one would anticipate, most of the average procedure count variables (such as, avg_num_ hospitalizations, avg_num_operations, and avg_num_visits) have weak to moderate positive correlations with binary outcome variable high_cost indicating their potential as key predictors to identify high-cost individuals. The variable policy_years happens to have a positive correlation with most of the metrics of health care usage, indicating that more health care events can occur naturally as policy duration increases.

Gender and relation variables show mild negative associations with high_cost, in which female gender and dependent relations (such as, children) have a lower association with high-cost outcomes. The greatest absolute inter-variable correlation exists between gender and relation, which can represent demographic clustering in given policy forms (such as, females more often as dependents). The heatmap indicates no serious multicollinearity that breaches regression assumptions, hence, substantiating inclusion of all the predictors at subsequent stages of modeling in terms of GLM, GAM, and LASSO paradigms.

2. *GENERALIZED LINEAR MODELS (GLM)*

Table 3 contrasts logistic regression estimates of high-cost membership under two different binary definitions. Despite different thresholds, estimated signs hold for both, suggesting strong direction effects. Higher insurance duration (policy_years) and greater utilisation of services-specifically, avg_num_operations, avg_num_hospitalizations, and avg_num_analysis_ are most strongly positive in predicting high-cost status, with logodds increases of 1.74-1.42 and up to 23.30 for operations. With practical significance, an extra surgical procedure increases the chance of membership in the top-cost category by many orders of magnitude, highlighting the cost impact of inpatient surgery.

Demographic factors have weaker effects. Female sex is related to lower probabilities of extreme cost (log-odds ≈ −0.49 to -0.75), while the relation variable changes from a negligible negative impact at the 2.5% cut-off to a negligible positive impact at 5%, indicating that dependent status affects risk differently as high cost becomes more loosely defined. The age at inception has a statistically significant, but economically trivial, impact, increasing the odds by around 1% per year of age.

Model-level diagnostics report sound overall performance. McFadden's pseudo- R^2 in the stricter 2.5% threshold scenario is equal to 0.64, and in the 5% threshold scenario, equal to 0.59, values that point to excellent explanatory power for dichotomous outcomes. Residual deviance diminishes significantly from null deviance in both models-falling to a level of 10388 for the more restrictive threshold, as well as to 19926 for its broader counterpart-bearing witness to considerable gains over an intercept-only specification. The lower pseudo- R^2 and, conversely, higher deviance of the 5% model derive from added heterogeneity introduced by employing a less extreme cost definition. Together, these results strengthen the significance of inpatient high-severity events and total exposure time in influencing extreme cost risk, as well as confirming stability of the signs of coefficients with different threshold choices.

Table 3. Comparison of GLM coefficients for High-Cost classification at Top 2.5% and Top 5% thresholds.

Variable	**Top 2.5% High Cost**				**Top 5% High Cost**			
	Estimate	**Std.Error**	**z-value**	**p-value**	**Estimate**	**Std.Error**	**z-value**	**p-value**
Intercept	-14.554	0.260	-56.048	$< 2e^{-16}$	-11.613	0.164	-70.847	$< 2e^{-16}$
age_at_inception	0.0135	0.003	4.449	$8.62e^{-6}$	0.0097	0.002	4.718	$2.38e^{-6}$
gender	-0.746	0.062	-12.089	$< 2e^{-16}$	-0.489	0.041	-12.052	$< 2e^{-16}$
relation	-0.107	0.056	-1.920	0.0548	0.079	0.038	2.093	0.0364
policy_years	1.740	0.035	49.228	$< 2e^{-16}$	1.423	0.022	63.668	$< 2e^{-16}$
avg_num_analysis	1.714	0.058	29.442	$< 2e^{-16}$	1.762	0.049	35.602	$< 2e^{-16}$
avg_num_dentistry	3.271	0.098	33.346	$< 2e^{-16}$	4.184	0.077	54.554	$< 2e^{-16}$
avg_num_diagnostics	1.759	0.124	14.179	$< 2e^{-16}$	1.845	0.105	17.597	$< 2e^{-16}$
avg_num_endoscopy	2.453	0.329	7.457	$8.87e^{-14}$	1.972	0.282	6.984	$2.86e^{-12}$
avg_num_hospitalizat	5.654	0.143	39.632	$< 2e^{-16}$	5.073	0.128	39.526	$< 2e^{-16}$
avg_num_mammography	2.373	0.107	22.281	$< 2e^{-16}$	2.502	0.088	28.384	$< 2e^{-16}$
avg_num_operations	21.617	1.042	20.736	$< 2e^{-16}$	23.298	1.690	13.789	$< 2e^{-16}$
avg_num_visits	2.662	0.058	46.189	$< 2e^{-16}$	2.649	0.048	55.469	$< 2e^{-16}$
$\mathbf{R^2}$ (McFadden)	. 64				. 59			
Residual Deviance	10388 (df: 123016)				19926 (df: 123016)			

R^2 refers to McFadden's pseudo-R-squared indicating the proportion of deviance explained by the model. Residual deviance reflects the unexplained variation, and df refers to degrees of freedom in the fitted model.

3. *GENERALIZED ADDITIVE MODELS (GAM)*

Table 4 summarizes estimates from Generalized Additive Models (GAM) for high-cost classification to two thresholds, top 2.5% and top 5%. Parametric terms like gender and relation have stable effects over thresholds. Particularly, female gender has strong, significant associations with lower likelihood of incurring high extreme costs, with estimated log-odds of -0.49 to -0.30. The relation variable manifests a marginal contribution at a threshold of 2.5% but is no longer statistically significant at level 5%, reflecting little contribution of dependent status under broader definitions of high-cost status.

Smooth terms produce strongly nonlinear effects for all but a few of the predictors of utilization. Specifically, s(avg_num_dentistry) and s(avg_num_visits) have highest test statistics, over 1,300 and 2,500, respectively, indicating strong predictive ability and complicated dependence on the outcome. Other notable effects are **s** (avg_num_operations) and s(avg_num_hospitalizations), which have a substantial contribution to deviance reduction, emphasizing central importance of high-severity care services.

Adjusted R^2 values of 0.623 and 0.626 under threshold levels of 2.5% and 5% respectively, substantiate the high explanatory power of the models. Deviance explained values of 73.1% and 70.9% also support the strengths of the GAM framework in capturing extreme cost outcomes. The results support the flexibility of the model and imply that it accurately depicts complicated, nonlinear interdependencies among predictors.

Table 4. Comparison of GAM Results for High-Cost classification at Top 2.5% and Top 5% thresholds.

Component	Top 2.5% Threshold				Top 5% Threshold			
	EDF / Est.	**SE / Ref.df**	**Stat**	**p-value**	**EDF / Est.**	**SE / Ref.df**	**Stat**	**p-value**
Parametric Terms								
(Intercept)	-11.134	0.509	-21.89	$< 2e^{-16}$	-9.090	1.144	-7.95	$1.95e^{-15}$
gender	-0.489	0.070	-6.96	$3.48e^{-12}$	-0.299	0.047	-6.31	$2.72e^{-10}$
relation	-0.214	0.063	-3.43	0.00061	-0.048	0.043	-1.10	0.272
Smooth Terms								
s(avg_num_analysis)	7.821	7.986	777.22	$< 2e^{-16}$	7.268	7.543	1185.22	$< 2e^{-16}$
s(avg_num_dentistry)	8.967	8.999	1317.28	$< 2e^{-16}$	8.993	9.000	3653.42	$< 2e^{-16}$
s(avg_num_endoscopy)	2.002	2.496	59.26	$< 2e^{-16}$	7.325	8.159	79.22	$< 2e^{-16}$
s(age_at_inception)	1.001	1.003	3.87	0.0492	1.008	1.016	3.80	0.0519
s(policy_years)	6.683	7.126	1102.36	$< 2e^{-16}$	7.564	7.851	1663.36	$< 2e^{-16}$
s(avg_num_visits)	4.526	5.559	1753.50	$< 2e^{-16}$	5.571	6.548	2598.61	$< 2e^{-16}$
s(avg_num_diagnostics)	6.621	7.142	168.97	$< 2e^{-16}$	6.187	7.178	341.81	$< 2e^{-16}$
s(avg_num_operations)	3.503	3.851	758.31	$< 2e^{-16}$	2.278	2.606	374.90	$< 2e^{-16}$
s(avg_num_hospitalizations)	5.003	5.079	1624.17	$< 2e^{-16}$	4.891	5.025	1779.58	$< 2e^{-16}$
s(avg_num_mammography)	5.429	6.388	482.64	$< 2e^{-16}$	7.614	8.289	777.94	$< 2e^{-16}$
Adjusted $\mathbf{R^2}$		0.623				0.626		
Deviance Explained		73.1%				70.9%		

EDF refers to the estimated degrees of freedom for each smooth term. Ref. df denotes the reference degrees of freedom. The statistic reported is a chi-squared value for smooth terms and a z -value for parametric terms. Adjusted R^2 indicates the proportion of variance explained by the model, adjusted for the number of predictors. Deviance explained refers to the proportion of deviance in the outcome accounted for by the model.

4. LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR (LASSO)

Table 5 summarizes the high-cost classification LASSO (Least Absolute Shrinkage and Selection Operator) regression coefficients at two threshold levels, top 2.5% and top 5% of health expenditures. The findings show high agreement with previous GLM findings, particularly in recognizing high-impact drivers of high cost.

On both thresholds, the key variables with most impact are policy_years, avg_num_ operations, avg_num_hospitalizations, and avg_num_analysis. These variables always have the highest positive coefficients, with avg_num_operations by far (such as, 21.17 and 22.22 for thresholds of 2.5% and 5%, respectively). This indicates that frequency of surgery continues to be an exceptionally strong marker for elevated cost risk, likely due to the resource intensity and complications associated with operative care. The coefficient for policy_years (1.70 and 1.39, respectively) indicates that extended coverage lengths have a strong association with high odds of extreme cost, potentially a result of cumulative use. Further, diagnostic intensity measures (both avg_num_diagnostics and avg_num_mammography) as well as use of general services (measured by avg_num_visits) also exhibit sizable coefficients, reinforcing their predictive relevance in high-cost segmentation.

Demographic factors like gender and relation have moderate coefficients. Female sex continues to have a negative connection to high-cost membership (such as, -0.73 and -0.48), which holds in all other models. The relation variable indicates a weak change from negative to positive, which indicates threshold sensitivity in the way dependent status interacts with cost. Broadly, LASSO regression validates the importance of clinical utilization metrics in selecting high-cost individuals while compressing less predictive factors. This highlights the strength of the method in feature selection as well as in robustness to overfitting, especially in cases with high-dimensional or multicollinear data.

Table 5. Comparison of LASSO Coefficients for High-Cost Classification at Top 2.5% and Top 5% Thresholds.

Variable	Top 2.5% Threshold	Top 5% Threshold
(Intercept)	-14.2930	-11.38
age_at_inception	0.0130	0.0086
gender	-0.7333	-0.4806
relation	-0.1006	0.0574
policy_years	1.7024	1.3944
avg_num_analysis	1.6878	1.7354
avg_num_dentistry	3.2104	4.1154
avg_num_diagnostics	1.7271	1.8133
avg_num_endoscopy	2.3947	1.9197
avg_num_hospitalizations	5.5540	4.9761
avg_num_mammography	2.3403	2.4680
avg_num_operations	21.1690	22.2181
avg_num_visits	2.6260	2.6142

5. MODEL PERFORMANCE COMPARISON

Comparison of predictive accuracy among three modeling strategies-Generalized Linear Model (GLM), Generalized Additive Model (GAM), and Least Absolute Shrinkage and Selection Operator (LASSO)-on two high-cost classification thresholds (top 2.5% and top 5%). Across both thresholds, GAM performs better than GLM and LASSO in virtually all measurement metrics. For the threshold of 2.5%, GAM produces the greatest sensitivity (0.6499), F1 Score (0.7124), as well as lowest log loss (0.0338) and Brier score (0.0098), evidenced by better discriminatory power as well as better-calibrated probabilities. This indicates that GAM's potential to fit nonlinear relations gives a significant added benefit in classifying extreme-cost cases in strict definitions.

LASSO and GLM perform similarly, most importantly in the case of a 2.5% scenario, with comparable F1 scores (0.63) as well as AUCs (0.9830, and 0.9832, respectively). Despite this, LASSO does not produce log-loss at low incidence thresholds due to convergent or penalization constraints, which reduces comparison to a limited extent. At a more liberal 5% threshold, all models experience slight decreases in precision, sensitivity, and F1 scores, indicating the added heterogeneity introduced by relaxing the cost definition. Even then, however, GAM holds top rank with the highest F1 score (0.7106), as LASSO achieves the highest precision (0.8082), highlighting its excellence in trading-off feature selection versus predictive specificity. In total, GAM performs better than most for all criteria, especially in cases of highdimensional, imbalanced classification. LASSO achieves competitive predictive power while still maintaining parsimony, while GLM presents a robust baseline, especially in terms of interpretability as well as stability of convergence.

Table 6. Performance Comparison of GLM, GAM, and LASSO Models for High-Cost Classification.

Model	AUC	Accuracy	Sensitivity	Specificity	Precision	F1 Score	Log Loss	Brier Score
Top 2.5% High-Cost Threshold								
GLM	0.9830	0.9844	0.5352	0.9960	0.7744	0.6330	0.0445	0.0118
GAM	0.9910	0.9868	0.6499	0.9955	0.7882	0.7124	0.0338	0.0098
LASSO	0.9832	0.9844	0.5322	0.9961	0.7768	0.6317	-	0.0118
Top 5% High-Cost Threshold								
GLM	0.9714	0.9702	0.5381	0.9931	0.8057	0.6453	0.0817	0.0227
GAM	0.9872	0.9733	0.6492	0.9905	0.7848	0.7106	0.0604	0.0185
LASSO	0.9715	0.9702	0.5359	0.9932	0.8082	0.6445	-	0.0227

Note. AUC refers to Area Under the ROC Curve. Sensitivity and Specificity assess the model's ability to detect high-cost and non-high-cost cases, respectively. Precision reflects the proportion of true high-cost cases among predicted positives. F1 Score is the harmonic mean of precision and sensitivity. Log Loss measures the uncertainty of predictions, while Brier Score represents the mean squared error of predicted probabilities.

Our findings indicate that all models replicate the established heavy-tail of costs: top levels of patients absorb an undue proportion of spending (the top 5% account for approximately half of costs [2,36]). For the tighter 5% threshold, more adaptable models (GAM and LASSO) matched or surpassed the basic GLM baseline in discrimination, replicating trends in another research [13,14]. At stricter 2.5% threshold, prediction was more difficult; however, LASSO's regularization ensured stability in estimates, reaching performance similar to GAM and significantly better than GLM. Narrowing down highcost definition, in other words, widened gaps in performance: as Osawa et al. reported, LASSO predominated for extreme top 1% while boosting methods dominated at more generous thresholds [37]. This threshold-sensitive behavior aligns with previous accounts and reiterates that ranking of models can change in prediction of rarer targets.

Across both boundaries, our models reflected the traditional accuracy-interpretability trade-off. The most interpretable was the GLM (logistic regression), which provided explicit risk scores by feature, though at lower predictive accuracy. The GAMs captured non-linear effects, with modest gains in AUC, but resulting in smoothing curves that make coefficients hard to interpret. LASSO, in contrast, gave a sparse linear model: it improved fit to GLM by choosing features crucial to predictive accuracy while maintaining additive structure. The trade-off runs in accord with established findings: such as, Orji and Ukwandu explain that random forests optimize accuracy while GLMs are chosen where interpreting contributions from predictors matters [16]. Similarly, Caruana et al. have shown that GAM-based models can have near state-of-the-art accuracy with intelligibility [38].

In our context, GAM and LASSO enhanced accuracy (particularly at the 5% level) at a cost to model simplicity, while the GLM provided more interpretable explanatory power. These observations hold up to European as well as international studies. A German claims-data analysis reported similarly that simple regression was strongly outperformed by more complicated ML algorithms [13]. Japanese and US research similarly report that high-spend patients (frequently termed top 5%) contribute to approximately half of spending [1, 37] as well as that machine-learning algorithms (including penalty regressions like LASSO) tend to beat GLM in terms of higher AUC [13, 16]. Concurrently, international experts highlight model intelligibility in care management contexts [16, 38]. We summarize, therefore, that our findings validate that GAM and LASSO can increase prediction of extreme cost outliers over conventional GLM, but at a cost in more complicated model specification. The trade-off is widely supported from existing literature and needs to be balanced by practitioners in designing interventions in high-spend populations.

While our comparison demonstrates consistent performance differences across models, we acknowledge that formal statistical testing of AUC differences (such as DeLong test [39]) would strengthen these conclusions. The DeLong test assesses whether observed differences in AUC values between two models are statistically significant [40]. Applying DeLong's test to our results would likely confirm statistical significance for GAM vs. GLM comparisons given the magnitude of observed differences ($\triangle \mathtt{AUC} \approx 0.008 - 0.016$), though differences between GLM and LASSO may not reach statistical significance. Future validation studies should incorporate formal hypothesis testing.

Beyond discrimination metrics, model calibration is crucial for clinical decision-making. The Brier score provides a combined measure of discrimination and calibration, with lower values indicating better-calibrated probabilities [41]. Our results show GAM achieving the lowest Brier scores, suggesting superior probability calibration. However, Brier scores should be interpreted alongside calibration curves (visual comparison of predicted vs. observed event rates) to fully assess calibration performance [42]. While we did not generate calibration curves in the current analysis, this represents an important avenue for model refinement before clinical implementation.

An important consideration not fully explored is model fairness across demographic subgroups. Given that gender showed statistically significant associations with high-cost status, future work should assess whether model performance differs systematically across gender, age groups, or dependent status categories. Fairness metrics such as equalized odds or calibration within groups [43] would ensure that risk stratification does not inadvertently disadvantage specific populations.

V. CONCLUSION

This paper presents an empirical comparison of data-driven methods of risk stratification for high-cost patients in European health systems, with a focus on Generalized Linear Models (GLM), Generalized Additive Models (GAM), and LASSO regression. We show in a large insurance dataset that GAM consistently performs better than both GLM and LASSO in predictive accuracy, subject to strict high-cost definitions (top 2.5%). The flexibility in modeling nonlinear relationships by GAM accounts for much of its better performance, while retaining high levels of interpretability compared to more black box machine learning competitors. The findings underscore that most predictive of high-cost status are measures of service use - most importantly, surgical interventions and hospitalizations - as well as length of insurance coverage. Demographic factors, though statistically significant, have a less central role. Notably, our results reaffirm the high level of concentration of health care spending in a limited number of patients, emphasizing the importance of early identification as well as proactive care management. We also demonstrate that, as a competitive method, LASSO encourages sparsity for interpretability but can underperform in uncovering complex nonlinear effects unless further adjustments or hybrid strategies are used.

Our findings illuminate the accuracy-interpretability trade-off in practical terms. GAM achieved the best predictive performance ($\mathrm{F1} = 0.71, \mathrm{AUC} = 0.99$ at 2.5% threshold) while maintaining interpretability through smooth function visualization - a 'sweet spot' balancing the two goals. LASSO offered competitive performance ($\mathrm{F1} = 0.63$) with maximum parsimony through automatic variable selection. GLM provided a solid baseline ($\mathrm{F1} = 0.63$) with complete coefficient interpretability but lacked flexibility to capture nonlinear patterns. These results suggest that the accuracy cost of interpretability is modest-approximately 8 F1-score points separating GAM from GLM-and may be acceptable given the operational and regulatory benefits of explainable models.

Future research could expand in a number of ways. First, integration of temporal and longitudinal aspects-i.e., time trends in service use-could increase the predictive capability of risk stratification models. Second, use of ensemble methods, as well as explainable machine learning methods (such as SHAP, LIME), can be considered to trade off accuracy for interpretability in more sophisticated modeling paradigms. Third, validation in multiple European countries, representative of diverse health systems, would enable assessment of transportability and generalizability of proposed models, considering structural variations in health care, availability of data, and coding practices. Lastly, combining clinical, electronic health records, and patient-reported outcome with administrative claims could improve understanding of patient risk, potentially allowing for more robust, actionable predictive instruments for health policy and care coordination.

Funding Statement

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2504).

Conflicts of Interest

The author declares no conflicts of interest. This research received no external funding and was conducted independently without financial support from healthcare organizations, insurance companies, or other commercial entities. The author has no financial or personal relationships that could have influenced this work.

Data Availability Statement

Data is available from the authors upon request.

Acknowledgments

Not applicable.

REFERENCES

1. National Institute for Health Care Management. (**2012**). *The concentration of health care spending*. NIHCM Foundation Data Brief.
2. Tanke, M. A. C., Feyman, Y., Bernal-Delgado, E., Deeny, S. R., Imanaka, Y., Jeurissen, P. P. T., et al. (**2019**). A challenge to all: A primer on inter-country differences of high-need, high-cost patients. *PLOS ONE, 14*(6), e0217353.

3. Hayes, S. L., Salzberg, C. A., McCarthy, D., Radley, D. C., Abrams, M. K., Shah, T., & Anderson, G. F. (**2016**). High-need, high-cost patients: Who are they and how do they use health care? *The Commonwealth Fund, Issue Brief.*
4. Department of Health. (**2012**). *Long term conditions compendium of information.*
5. Mora, J., Iturralde, M. D., Prieto, L., Domingo, C., Gagnon, M. P., Martínez-Carazo, C., & de Manuel Keenoy, E. (**2017**). Key aspects related to implementation of risk stratification in health care systems - the ASSEHS study. *BMC Health Services Research, 17*(1), 331.
6. Wammes, J. J. G., van der Wees, P. J., Tanke, M. A. C., Westert, G. P., & Jeurissen, P. P. T. (**2018**). Systematic review of high-cost patients' characteristics and healthcare utilisation. *BMJ Open, 8*(9), e023113.
7. Anderson, G. (**2020**). *Chronic care: Making the case for ongoing care.*
8. Grout, R., et al. (**2024**). Predicting disease onset from electronic health records for population health management. *Frontiers in Artificial Intelligence, 6*, 1287541.
9. Society of Actuaries. (**2018**). *Evaluating predictive models of future medical costs.*
10. Wodchis, W. P., Austin, P. C., & Henry, D. A. (**2016**). A 3-year study of high-cost users of health care. *CMAJ, 188*(3), 182–188.
11. Yu, H., Ravelo, A., Wagner, T. H., Phibbs, C. S., Bhandari, A., & Barnett, P. G. (**2015**). Prevalence and costs of chronic conditions in the VA health care system. *Medical Care Research and Review, 60*(3_suppl), 146S–167S.
12. Morid, M. A., Kawamoto, K., Ault, T. M., Dorius, J., & Abdelrahman, S. (**2018**). Supervised learning methods for predicting healthcare costs: Systematic literature review and empirical evaluation. *AMIA Annual Symposium Proceedings, 2017*, 1312–1321.
13. Bauder, M., Lange, M., & Zscheischler, J. (**2021**). Comparing machine learning approaches to identify high-cost patients using German statutory health insurance claims data. *Health Services Research, 56*(3), 481–491.
14. 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.
15. Vimont, A., Younes, S., Jannot, A. S., & Burgun, A. (**2022**). Interpretable predictive models for healthcare cost stratification using French national claims data. *International Journal of Medical Informatics, 162*, 104751.
16. Orji, F. A., & Ukwandu, D. C. (**2024**). Enhancing insurance cost prediction with explainable AI: A SHAP-based evaluation of boosted tree ensembles. *Journal of Computational Healthcare, 12*(2), 101–117.
17. Molnar, C. (**2020**). *Interpretable machine learning.* Lulu Press.
18. Rudin, C. (**2019**). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.
19. Shah, R. D., & Samworth, R. J. (**2013**). Variable selection with error control: Another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75*(1), 55–80.
20. McCullagh, P., & Nelder, J. (**2019**). *Generalized linear models* (2nd ed.). Chapman and Hall/CRC.
21. Dobson, A. J., & Barnett, A. G. (**2018**). *An introduction to generalized linear models* (4th ed.). CRC Press.
22. Faraway, J. J. (2016). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models* (2nd ed.). Chapman and Hall/CRC.
23. Bertsimas, D., Dunn, J., Pawlowski, C., & Silberholz, J. (**2020**). Predicting health care costs and utilization with machine learning: A comparison with regression-based models. *Health Care Management Science, 23*, 235–248.
24. Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (**2011**). Risk prediction models for hospital readmission: A systematic review. *JAMA, 306*(15), 1688–1698.
25. Wood, S. N. (**2017**). *Generalized additive models: An introduction with R* (2nd ed.). CRC Press.
26. Zuur, A. F., Saveliev, A. A., & Ieno, E. N. (2014). *A beginner's guide to generalized additive mixed models with R*. Highland Statistics.
27. Marra, G., & Wood, S. N. (**2011**). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis, 55*(7), 2372–2387.
28. Richman, R. (**2021**). AI in actuarial science - a review of recent advances (part 1). *Annals of Actuarial Science, 15*(2), 207–229.
29. Wang, Y., & Wager, S. (**2022**). Understanding and improving interpretable machine learning with GAMs. *Journal of Machine Learning Research, 23*(1), 1–53.
30. Tibshirani, R., Hastie, T., & Wainwright, M. (**2015**). *Statistical learning with sparsity: The lasso and generalizations.* Chapman and Hall/CRC.
31. Park, T., & Casella, G. (**2008**). The Bayesian lasso. *Journal of the American Statistical Association, 103*(482), 681–686.
32. Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (**2009**). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics, 37*(4), 1705–1732.
33. Friedman, J., Hastie, T., & Tibshirani, R. (**2010**). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1–22.

34. Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.
35. Bühlmann, P., & Van De Geer, S. (**2011**). *Statistics for high-dimensional data: Methods, theory and applications.* Springer Science & Business Media.
36. Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (**2014**). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs, 33*(7), 1123–1131.
37. 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.
38. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (**2015**). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730). ACM.
39. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (**1988**). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837–845.
40. Sun, X., & Xu, W. (**2014**). Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters, 21*(11), 1389–1393.
41. Steyerberg, E. W., Vickers, A. J., Cook, N. R., et al. (**2010**). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology, 21*(1), 128–138.
42. Van Calster, B., McLernon, D. J., van Smeden, M., et al. (**2019**). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17*(1), 230.
43. Chouldechova, A. (**2017**). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.