Healthcare Analytics 8 (2025) 100431

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

A log-linear analytics approach to cost model regularization for inpatient stays through diagnostic code merging

Chi-Ken Lu *, David Alonge, Nicole Richardson, Bruno Richard

Department of Mathematics and Computer Science, Rutgers University, Newark, 07302, NJ, USA

ARTICLE INFO

Keywords:
High-dimensional regression
Coefficient consistency
Implicit regularization
Variables grouping
Hessian matrix
Cost prediction

ABSTRACT

Healthcare cost models that use a great number of detailed ICD-10 diagnostic codes produce unstable results, yet the underlying causes of this instability have not been well understood. This study provides a mathematical framework linking the variability of model coefficients to the uneven, power-law distribution of diagnostic codes and the structure of the regression model. We propose a transparent approach that improves coefficient stability by merging similar codes through hierarchical truncation. Using Medicare data, we demonstrate how this method clarifies the trade-off between code detail and model reliability, offering analysts and policymakers a practical and interpretable tool for diagnosis-based cost modeling.

1. Introduction

Accurate and interpretable cost models are essential in healthcare research as they provide critical tools for estimating, analyzing, and understanding healthcare spending patterns [1–4]. Linear and log-linear regression methods are a particularly popular approach to modeling healthcare data as they are relatively simple to implement and highly interpretable [5–7]. Log-linear models can easily leverage individual-level features that are prevalent in healthcare datasets (e.g., demographic and diagnostic information) to predict outcomes such as the cost of care for acute and chronic conditions, and account for the characteristics of cost distributions [8–11]. A prominent example is the Center for Medicare and Medicaid Services (CMS), which uses regression-based models in its risk adjustment methodologies for healthcare payment systems [12,13]. These models produce regression coefficients that form the basis for risk scores to determine reimbursements for enrollees with diverse demographics and health statuses, which makes their accuracy crucial for both the patient and the provider.

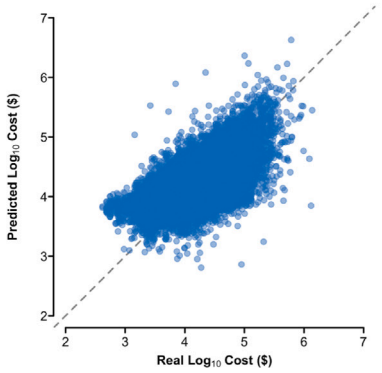
Like many other domains, healthcare data is inherently high-dimensional, consisting of a large number of patient observations (n) measured across numerous features (p), such as demographics [14], provider information, financial records, and diagnostic codes. This high dimensionality poses challenges for learning stable and robust representations, often necessitating the use of regularization or dimensionality reduction methods. For example, Ridge regression introduces a coefficient-shrinking penalty that reduces variance and mitigates overfitting, albeit at the cost of additional bias [15,16]. In deep learning, pooling layers similarly reduce feature dimensionality, enabling more efficient and meaningful representation learning [17]. However, these nonlinear operations, while improving flexibility, often reduce interpretability by obscuring how individual variables influence predictions. These challenges are especially pronounced in estimating the cost of inpatient stays, which represent a substantial share of U.S. healthcare expenditures [18]. Because each stay is linked to numerous diagnostic variables, typically encoded using the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) [19,20], achieving accurate and consistent cost prediction remains a critical problem at the intersection of machine learning and healthcare.

Problem statement. This study analyzes inpatient cost data from the New York Downstate subset of the Medicare Provider Analysis and Review (MedPAR) dataset. Each inpatient stay has at most 25 assigned diagnosis codes. The outcome variable y is the log-transformed cost of the stay, and the predictors $\mathbf{x}$ are binary variables indicating the presence of specific ICD-10 codes. We apply OLS regression to fit the model across randomized training subsets ($n_{\text{tr}} \approx 400{,}000$, $p \approx 20{,}000$), evaluating on a holdout test set ($n_{\text{te}} \approx 100{,}000$). The average training R^2 is approximately 0.45, with a test R^2 near 0.41. Panel A in Fig. 1 displays the predicted log cost for test data against its true value. Despite the predictive scores being better than those reported for HHS-HCC risk adjustment models [13], the OLS regression coefficients are highly inconsistent across subsamples. Panel B in Fig. 1 lists a few regression coefficients from using different training data. This instability makes it difficult to use these coefficients for developing a reliable ICD-10-based risk score.

* Corresponding author.
E-mail address: CL1178@rutgers.edu (C.-K. Lu).

https://doi.org/10.1016/j.health.2025.100431
Received 23 August 2025; Received in revised form 3 November 2025; Accepted 4 November 2025
Available online 6 November 2025

A



B

	Set 1	Set 2	Set 3
M4144	2.163	1.826	2.281
I82452	1.682	1.712	1.663
H18421	1.526	0.304	0.000
S13153A	-1.987	0.267	-1.889
T3176	-2.067	-1.167	-1.530

Fig. 1. Panel A: the predictive log costs from OLS models against their true values. The OLS models include explainable variables: indicators of ICD-10 diagnostic codes and demographics (age, sex, and race). Panel B: the inconsistency among a few regression coefficients from OLS fitting to different training data. In Set 3, the corresponding training samples do not contain H18.421 so that the fitted coefficient is zero.

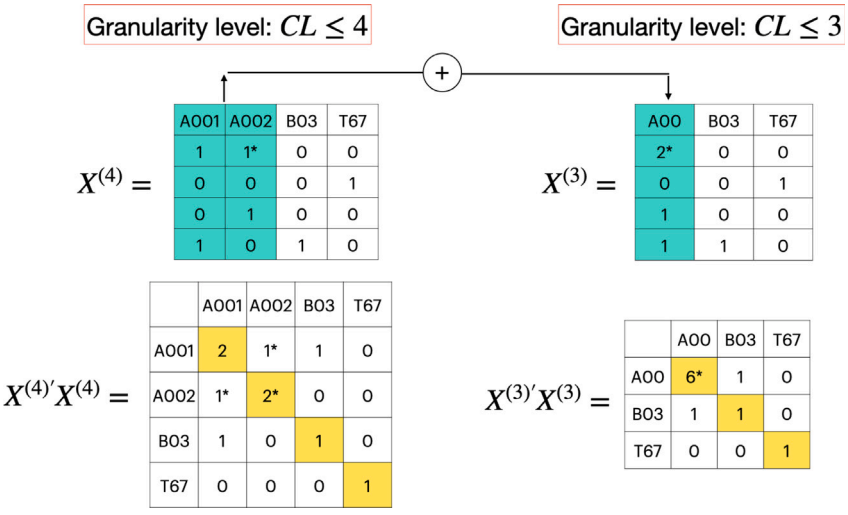


Fig. 2. A toy example for illustration of reducing code granularity by merging similar codes. The design matrix $X^{(4)}$ on the left records the diagnoses using codes with $CL \leq 4$. Then the granularity is lowered by truncating codes with four characters to 3. The predictors for A001 and A002 are added to form the new predictor A00 in the matrix $X^{(3)}$ on the right. The corresponding Hessian matrices are displayed at the bottom. Lowering granularity (left to right) increases the trace due to the co-occurrence of merged codes in the same stay. If the marked 1 in $X^{(4)}$ is set to 0, then the rest of the marked numbers change and the traces on both sides become identical.

Proposed solution. We address OLS coefficient instability in diagnosis-based cost models by introducing an implicit regularization mechanism through ICD-10 code merging via truncation and aggregation (Fig. 2). This approach reduces dimensionality while preserving the hierarchical structure of diagnostic codes. The need for such regularization arises because ICD-10 code frequencies follow a power-law distribution (Fig. 4-F), where a few codes occur extremely often while most are rare, producing small Hessian eigenvalues and large coefficient variances. Code truncation mitigates this imbalance by pooling infrequent codes, which increases the Hessian trace (Lemma 2) and improves coefficient stability. Although truncation has been used to address sparsity in disease prediction [21–23], its role in stabilizing coefficients and its theoretical link to the Hessian structure have not been examined. To measure this effect, we propose a stability metric η (defined in Eq. (11)) based on the Spearman correlation between coefficients across subsamples, and compare results against Ridge regression and DRG/HCC groupings as empirical benchmarks.

The paper is organized as follows. Section 2 reviews prior studies on modeling inpatient costs using diagnosis codes. Section 3 presents a descriptive analysis of the data subset, examines the structure of the design and Hessian matrices associated with ICD-10 codes, and discusses variable groupings by code truncation as an implicit regularization method for OLS. It also explores the variance of OLS coefficient estimates in relation to the Hessian matrix, introduces a metric for measuring coefficient consistency, and analyzes Hessian matrix characteristics for the dataset. Section 4 reports coefficient consistency – an equally important factor as predictive accuracy in cost models – across varying levels of ICD-10 code granularity. Section 5 offers insights into the OLS modeling results under code truncation and discusses limitations. Finally, Section 6 summarizes the findings and outlines future research directions.

2. Related work

Linear regression has long served as a cornerstone in healthcare cost modeling due to its simplicity and interpretability [1,9,24]. However, diagnosis-based models often involve thousands of ICD-derived predictors, creating sparsity and multicollinearity. Prior studies have mitigated this issue through dimension reduction or diagnostic code grouping. For example, [21,23] used truncated ICD codes to stabilize

sparse models, while [9] grouped ICD-10 codes into diagnostic categories alongside demographic variables. Grouping systems such as Diagnosis-Related Groups (DRGs) [25,26] and Hierarchical Condition Categories (HCCs) [27,28] remain central to risk-adjustment frameworks like those implemented by CMS [13]. Coefficient consistency has also been examined in dynamic cost models that track changing patient status over time [29,30]. To our knowledge, no prior study in healthcare analytics has formally linked OLS coefficient instability to the Hessian matrix structure or demonstrated code truncation as a form of implicit regularization.

Despite their prevalence, OLS models are sensitive to right-skewed and heteroscedastic cost data [9,10]. Log-transformations can mitigate these effects [31], but high-dimensional and unevenly distributed predictors still yield unstable coefficients. Explicit regularization methods such as ridge [15] and Lasso [32–34] alleviate multicollinearity, yet they do not address instability arising from the structural granularity of diagnostic codes.

Recent work has increasingly adopted machine learning (ML) models for cost prediction [6,7,10,21,23,35,36]. Studies have used random forests and gradient boosting for claims data in New York [11], Texas [37], and Louisiana [38], generally finding superior predictive accuracy compared with linear models [9]. Tree-based and deep learning models capture nonlinear dependencies but at the cost of interpretability. Although implicit regularization mechanisms such as early stopping and stochastic gradient descent [17,39,40] help prevent overfitting, these models remain analytically opaque—limiting their utility for policy interpretation or coefficient-level inference.

By contrast, linear models retain the advantage of established variance analysis [41–43], allowing investigation into coefficient consistency—a topic rarely explored in healthcare cost modeling. To our knowledge, no prior work formally links OLS coefficient variance to the structural properties of diagnostic codes. The present study addresses this gap by developing a log-linear framework that connects the Hessian matrix to coefficient instability and by demonstrating that ICD-10 code truncation and variable aggregation act as an implicit regularization mechanism that improves both coefficient stability and interpretability.

3. Methods

This study develops a diagnosis-based log-linear regression framework for modeling inpatient costs using ICD-10 diagnostic codes. A central methodological challenge is the instability of OLS coefficients when models are refitted on random subsamples, a phenomenon arising from the heavy-tailed, power-law distribution of ICD-10 code frequencies. To quantify this effect, we define a coefficient stability metric and show that small eigenvalues of the Hessian matrix are the primary source of high variance. Two complementary stabilization strategies are examined: (i) Ridge regression, which inflates small eigenvalues via an L_2 penalty, and (ii) hierarchical code merging through ICD-10 truncation, which increases the Hessian trace and serves as an implicit regularization mechanism. Unlike conventional grouping schemes such as DRG or HCC, the proposed approach maintains diagnostic granularity while enhancing coefficient stability in a theoretically grounded manner. Demographic variables are excluded from the theoretical analysis, as their relatively uniform distributions do not contribute to the observed instability.

3.1. Data

The Medicare Provider Analysis and Review (MedPAR) Limited Data Set contains comprehensive, discharge-level information for all Medicare and Medicaid beneficiaries who receive inpatient hospital services in the United States. Each record represents a single inpatient stay, aggregating all claims associated with a continuous hospitalization episode—from admission to discharge. The dataset includes detailed variables describing patient demographics, hospital identifiers, financial information (e.g., total charges and payments), diagnostic and procedural codes, and time-related elements such as admission and discharge dates.

For this study, we analyze a geographically defined subset of the FY2018 MedPAR file. To maintain a manageable data scale while preserving adequate sample size and clinical diversity, we restrict the sample to Downstate New York, encompassing Westchester, Bronx, New York (Manhattan), Queens, Kings (Brooklyn), Richmond (Staten Island), Nassau, and Suffolk counties (Fig. 3A). This regional subset constitutes roughly 3% of the national MedPAR dataset per fiscal year.

While the Downstate New York sample remains broadly comparable to the full U.S. MedPAR population, several differences are noteworthy. Hospitals in this region are generally larger and report higher mean inpatient costs, consistent with known regional cost-of-living and case-mix differences. As illustrated in Fig. 3B, average inpatient costs in Downstate New York exceed those observed both in Upstate New York and nationally. Consequently, while the proposed modeling framework is data-agnostic and generalizable, the estimated OLS coefficients from this regional sample may not be directly representative of the broader U.S. population.

Cost variation across stays also correlates with the number of assigned diagnosis codes (Fig. 3C), reflecting greater clinical complexity among higher-cost cases. In terms of demographics (Fig. 3D), Downstate New York exhibits greater racial and ethnic diversity – with relatively fewer White patients and higher proportions of Black, Asian, and Hispanic beneficiaries – while age and sex distributions remain consistent with national MedPAR data.

3.2. Ordinary least square and binary variables for ICD codes

Let $y = \log_{10}(\text{cost})$ represent the logarithm (base 10) of the inpatient stay cost. This quantity is modeled as a linear function of p binary predictors $x_1, x_2, \ldots, x_p$, which indicate the presence of corresponding ICD-10 diagnosis codes $\alpha_1, \alpha_2, \ldots, \alpha_p$. Up to a zero-mean noise term ϵ, the model takes the form

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \epsilon, \tag{1}$$

where $\beta = \beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients to be estimated. The coefficients are obtained by minimizing the OLS objective function:

$$\mathcal{L}_{\text{OLS}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{2}$$

where n is the number of observations. The closed-form OLS solution is given by

$$\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y, \tag{3}$$

where $\tilde{X}$ is the design matrix augmented with an intercept column

$$\tilde{X} = \begin{bmatrix} X_{(n\times p)} & \mathbf{1}_{(n\times 1)} \end{bmatrix}, \tag{4}$$

The corresponding $(p+1)\times(p+1)$ Hessian matrix of the squared loss objective is

$$\tilde{X}'\tilde{X} = \begin{bmatrix} X'X & X'\mathbf{1} \\ \mathbf{1}'X & n \end{bmatrix}. \tag{5}$$

The variance–covariance matrix of the estimated coefficients is $\hat{\sigma}^2(\tilde{X}'\tilde{X})^{-1}$ where $\hat{\sigma}^2$ is the estimated variance of the noise term ϵ. The variance of an individual coefficient $\hat{\beta}_i$ is therefore

$$\text{Var}(\hat{\beta}_i) = \hat{\sigma}^2 v_i, \tag{6}$$

with v_i being the ith diagonal element of $(\tilde{X}'\tilde{X})^{-1}$ [16]. This term directly reflects the sensitivity of coefficient estimates to the structure and sparsity of the design matrix—a key factor in the inconsistency of regression coefficients in high-dimensional OLS models.

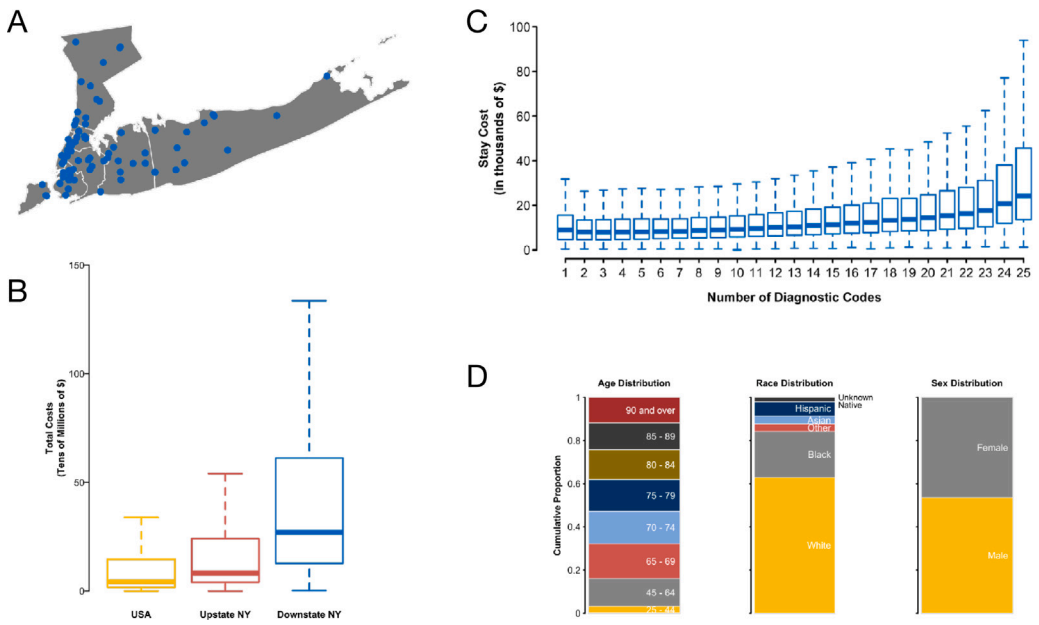


Fig. 3. **A**. Map of the hospitals that make up the downstate New York subset we use in our analysis. **B**. Distribution of total costs for hospitals in the greater USA, Upstate New York, and downstate New York in the FY2018 subset. **C**. Association between the average cost of a stay and the number of diagnostic codes attached to the stay for hospitals in Downstate New York. The more diagnostic codes are attached to a stay, the more expensive the stay. **D**. Distributions over the age, sex, and race variables in the MedPAR subset.

Statistically, a large value of v_i indicates that the estimate $\hat{\beta}_i$ will vary significantly when different subsets of the data are used to fit the model. Unfortunately, the diagonal entries of an inverse matrix are highly nonlinear functions of the entries in the original matrix, which makes it challenging to control the variance directly. To mitigate this, one can explicitly add a penalty term to the loss function – such as in Ridge regression – which increases the eigenvalues of the Hessian and thereby reduces all variances v_i. We present the following remark, which provides an important observation regarding the total variance of the coefficients and its relationship to the eigenstructure of the design matrix.

Remark 1. Sum of coefficient variance $S_V = v_1 + v_2 + \cdots + v_p + v_0$, sum of eigenvalues of inverted Hessian $S_I = s_1^{-1} + s_2^{-1} + \cdots s_p^{-1} + s_0^{-1}$, and trace of Hessian satisfy the following relations,

$$S_V = S_I > \frac{1}{tr(\tilde{X}'\tilde{X})} . \tag{7}$$

Proof. Let the set $\{s_1, s_2, \ldots, s_p, s_0\}$ denote the eigenvalues of the Hessian $\tilde{X}'\tilde{X}$. It follows that the inverse, $(\tilde{X}'\tilde{X})^{-1}$, has the set of eigenvalues $\{s_1^{-1}, s_2^{-1}, \ldots, s_p^{-1}, s_0^{-1}\}$. Using the fact that the sum of eigenvalues of a square matrix is identical to the sum of its diagonal entries, then the equality follows as the square matrix is the inverse of Hessian. Since the Hessian is positive definite, all eigenvalues are positive. Since $f(s) = 1/s$ is convex for $s > 0$, we can use Jensen's inequality to show $S_I = \mathbf{E}[1/s] > 1/\mathbf{E}[s]$. Note that the sums have been replaced with an expectation. Finally, $s_1 + s_2 + \cdots + s_p + s_0 = \mathbf{E}[s] = tr(\tilde{X}'\tilde{X})$, which completes the proof.

This highlights that both the eigenvalues and the diagonal elements of the Hessian matrix play a critical role in determining the stability of coefficient estimates and the model's predictive performance. Therefore, understanding what structural properties of the design matrix contribute to large variance v values is essential.

We now examine the structure of the matrix $X'X$, which is the main component of the block matrix in Eq. (5). Define the frequency F_j of ICD code j as the column sum $F_j := \sum_i X_{ij}$, and define D_i as the number of diagnoses recorded for stay i, i.e., the row sum $D_i := \sum_j X_{ij}$. Due to MedPAR formatting, $D_i \leq 25$, which results in a highly sparse matrix X.

Remark 2. The $p \times p$ matrix $X'X$ is, in fact, the co-occurrence matrix for the ICD codes. That is the jth diagonal entry $[X'X]_{jj} = F_j$, the total frequency of the ICD code j. The off-diagonal entries $[X'X]_{jk}$ represent the co-occurrence of the ICD code j and code k. Consequently, the sum of eigenvalues $\sum_{i=1}^{p} s_i = \sum_{i=1}^{p} F_i$.

Proof. The diagonal entry $[X'X]_{jj} = \sum_i (X_{ij})^2$, the L^2-norm of the binary column $X_{:j}$. As 0 and 1 are invariant under square, $[X'X]_{jj} = \sum_i X_{ij} = F_j$. Similarly, off-diagonal entry $[X'X]_{jk} = \sum_i X_{ij}X_{ik}$, a count of stays having with ICD-10 codes j and k simultaneously in the record. As the diagonal entries are connected with the ICD code frequencies, using the property of trace can show that the sum of ICD-10 code frequencies in the data equal the sum of eigenvalues.

For example, in the FY-2018 MedPAR subset, there are approximately 500,000 stays and over 20,000 unique ICD-10-CM codes. The frequencies of individual codes vary widely, ranging from 1 to over 300,000. The average frequency per stay per code, computed as $\mathrm{tr}(X'X)/(np)$, is approximately 0.00073, reflecting the sparsity of the data.

In the following sections, we describe both explicit and implicit techniques for increasing the trace of the Hessian matrix. These techniques effectively reduce the variance of the estimated regression coefficients, thereby improving the consistency and reliability of the log-cost model.

3.3. Explicit regularization and effective dimension

By adding an $L2$-norm penalty $\lambda \sum \beta_i^2$ to the original objective $\mathcal{L}_{OLS}$, we obtain the Ridge regression estimator

$$\hat{\beta}_{Ridge} = (\tilde{X}'\tilde{X} + \lambda I_{p+1})^{-1}\tilde{X}'\mathrm{y} , \tag{8}$$

where $\lambda > 0$ is the regularization parameter, and I_{p+1} is the identity matrix of size $(p+1)\times(p+1)$. The added penalty term shrinks the coefficients toward zero, resulting in estimates typically closer to zero than

those from OLS. The lower bound for the sum of coefficient variances in Eq. (7) is also affected by regularization, becoming: $1/[tr(\tilde{X}'\tilde{X}) + (p+1)\lambda]$. Thus, increasing λ decreases the total variance of the estimated coefficients $\hat{\beta}$, enhancing their stability.

The expected prediction error (risk) $\mathbb{E}[(y-\mathbf{x}'\hat{\beta})^2]$ can be decomposed into three components: the irreducible noise, squared bias, and variance [16]. Regularization reduces variance at the cost of increased bias since the penalty term pulls estimates away from their true values. In effect, this reduces model complexity. The effective dimension of the Ridge model, which quantifies the model's complexity, is given by the trace of the matrix S that maps the observed responses to the predicted ones: $\hat{\mathbf{y}} = S\mathbf{y}$ [16,44],

$$\rho = tr[\tilde{X}(\tilde{X}'\tilde{X} + \lambda I)^{-1}\tilde{X}'] = \sum_{i=0}^{p} \frac{s_i}{s_i + \lambda}, \tag{9}$$

where s_i are the eigenvalues of $\tilde{X}'\tilde{X}$.

Although ρ provides insight into model capacity, computing it exactly requires knowledge of all eigenvalues, which can be computationally expensive. However, the function $f(s) = \frac{s}{s+\lambda}$ is concave, allowing for a convenient upper bound.

Lemma 1. *The effective dimension for the Ridge regression model trained on n data has a upper bound ρ_B,*

$$\rho \le \rho_B = \frac{(p+1)\bar{s}}{\bar{s} + \lambda/n}. \tag{10}$$

where $\bar{s}$ stands for the average of the eigenvalues of $\tilde{X}'\tilde{X}/n$.

Proof. As the number of predictors $p \gg 1$, the effective dimension can be evaluated as an integral, $\rho \approx (p+1)\int ds \frac{s}{s+\lambda} p(s) = (p+1)\mathbf{E}[\frac{s}{s+\lambda}]$ where $p(s)$ stands for the density of eigenvalues of Hessian matrix. Since the function $f(s) = \frac{s}{s+\lambda}$ is concave, i.e. $f''(s) < 0$ for all $s > 0$ provided $\lambda > 0$. The above inequality directly follows Jensen's inequality.

This upper bound for effective dimension ρ_B (Eq. (10)) is much easier to compute and is useful for analyzing and comparing predictive performance across different levels of regularization.

3.4. Consistency metric

We repeatedly split the data into training and test sets to assess consistency, estimating regression coefficients β each time. If the estimates are consistent, then coefficient pairs $(\beta^{(a)}, \beta^{(b)})$ from different splits should be relatively equal. We treat these pairs as samples from a bivariate distribution and measure their agreement using Spearman correlation, which is less sensitive to outliers. We define the consistency metric η as the mean Spearman correlation over all distinct pairs

$$\eta = \frac{1}{N(N-1)} \sum_{a \ne b} r_s(\beta^{(a)}, \beta^{(b)}) \tag{11}$$

Higher values of η (close to 1) indicate strong consistency, while lower values suggest instability, with coefficient scatter resembling a circle.

3.5. Varying code granularity as implicit regularization

ICD-10-CM codes have character lengths of 7, beginning with an uppercase letter (A-Z), indicating the broad disease category, followed by two numbers that indicate the disease within the broad category. This is followed by 3–4 characters. To maintain semantic meaning, the merging of ICD-10 codes can therefore have a code length (CL) of 3 to 7 (i.e., $3 \le CL \le 7$). In our dataset, the set of codes α contains approximately $p = |\alpha| \approx 20{,}000$ unique ICD-10 codes (see Section 3.2). To reduce the number of predictors p, we propose shortening longer codes by truncating them to a fixed length l, merging similar codes, and reducing dimensionality. Specifically, for any code with $CL > l$, we retain only its first l characters. In this study, we explore the impact of truncating ICD-10 codes to have l between 2 and 7 characters, whereby 7 is the full ICD-10 diagnostic code. Reduced sets after truncation are denoted by $\alpha^{(l)}$, with $p^{(l)} = |\alpha^{(l)}|$. We define the truncation function

$$T^{(CL\le l)}(\text{code}) = \text{code}[0:l], \tag{12}$$

using Python-style string slicing. The granularity of the codes is controlled by l; decreasing l reduces the number of unique codes $p^{(l)}$. For instance, $T^{(CL\le 6)}$(T670XXA) = $T^{(CL\le 6)}$(T670XXD) = T670XX, effectively merging both codes into a single one in $\alpha^{(6)}$. The progression $\alpha \to \alpha^{(6)} \to \cdots \to \alpha^{(2)}$ is analogous to hierarchical reductions used in ICD-9 tree structures [22].

Next, we describe how the predictors x_i and matrix X evolve under this dimensionality reduction. Drawing an analogy from convolutional neural networks (CNNs) [17,45,46], where pooling layers aggregate pixels into coarser representations, we define a similar "sum-pooling" operation for variables. Specifically, the coarser variable $x_i^{(l)}$ is computed as a sum of finer variables

$$x_i^{(l)} = \sum_j Q_{ji}^{(m\to l)} x_j^{(m)}, \tag{13}$$

where $m > l$, and the matrix $Q^{(m\to l)} \in \mathbb{R}^{p^{(m)}\times p^{(l)}}$ has entries $Q_{ji} = 1$ if the truncated code $T^{(CL\le l)}(\alpha_j^{(m)}) = \alpha_i^{(l)}$, and 0 otherwise. This operation merges columns in $X^{(m)}$ to produce the coarser matrix $X^{(l)}$, preserving the total number of diagnoses per stay (i.e., row sums remain unchanged). The association of cost and number of codes attached to stay (Fig. 3-**C**) is retained. Fig. 2 illustrates this process using a toy matrix with $CL \le 4$, reduced to $CL \le 3$ by merging columns. The regression coefficients $\beta^{(l)}$ corresponding to $\alpha^{(l)}$ are estimated via

$$y = \beta_0^{(l)} + \sum_{i=1}^{p^{(l)}} \beta_i^{(l)} x_i^{(l)} + \epsilon, \tag{14}$$

with OLS still applicable. However, we now use the transformed design matrix

$$X^{(l)} = X^{(m)} Q^{(m\to l)}, \tag{15}$$

replacing the one in Eq. (4). The tilded matrix $\tilde{X}^{(l)} = [X^{(l)}, \mathbf{1}]$ maintains the same number of rows but has fewer columns than $\tilde{X}^{(l+1)}$, reflecting the reduced granularity. Importantly, reducing granularity increases the trace of the Hessian matrix, which influences the variance of regression coefficients:

Lemma 2. *The trace of the Hessian matrix increases as granularity decreases*

$$tr(X^{(l)\prime}X^{(l)}) \ge tr(X^{(l+1)\prime}X^{(l+1)}), \tag{16}$$

with equality only if codes in $\alpha^{(l+1)}$ that map to the same code in $\alpha^{(l)}$ never co-occur.

Proof. The proof essentially bases on the observation that two non-negative columns have $(x_1 + x_2)\cdot(x_1 + x_2) \ge x_1\cdot x_1 + x_2\cdot x_2$ with equality holds only if $x_1 \cdot x_2 = 0$. Use the trace formula, $tr(X^{(l)\prime}X^{(l)}) = \sum_{jkm} X_{jk}^{(l+1)} X_{jm}^{(l+1)} \sum_i Q_{ki}Q_{mi}$. For ease of notation, we have dropped the superscript in Q. The latter sum results in the entry $(QQ')_{kl}$, representing whether the more granular code $\alpha_k^{(l+1)}$ and code $\alpha_m^{(l+1)}$ are combined into the same code in the less granular code set $\alpha^{(l)}$. Then the matrix QQ' is sparse and the diagonal of it is all ones. Thus, $tr(X^{(l)\prime}X^{(l)}) = \sum_{jkm} X_{jk}^{(l+1)} X_{jm}^{(l+1)}[I_{km} + (QQ' - I)_{km}]$ from which we conclude that $tr(X^{(l)\prime}X^{(l)}) \ge tr(X^{(l+1)\prime}X^{(l+1)})$ and the equality holds when the contribution $\sum_j X_{jk}^{(l+1)} X_{jm}^{(l+1)}$ from any combined pair (k,m) is zero.

This is illustrated in Fig. 2, where reducing from $CL \le 4$ to $CL \le 3$ increases the trace from 6 to 8. If codes A001 and A002 did not co-occur in the same row, the trace would remain unchanged during reduction.

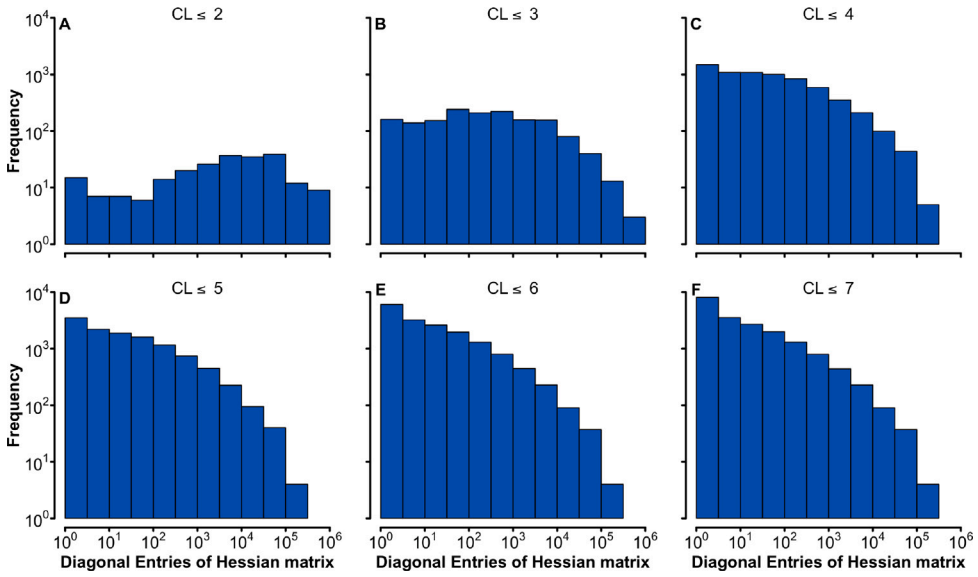


Fig. 4. Evolving histograms of diagonal entries of Hessian matrix for different code granularity $CL \le l = [2, 3, 4, 5, 6, 7]$ (panel A to F). In panels D-F for higher granularity, the diagonal entries display power law distributions.

Table 1
Data characteristics: the number of predictors $p^{(l)}$ and the mean eigenvalue $\bar{s}^{(l)}$ of scaled Hessian matrix $X'X/n$ with $n = 495\,023$ observation in FY2018 subset. The ICD code granularity is specified by the character length's upper limit, $CL \le l$.

$CL \le 2$		$CL \le 3$		$CL \le 4$		$CL \le 5$		$CL \le 6$		$CL \le 7$	
$p^{(2)}$	$\bar{s}^{(2)}$	$p^{(3)}$	$\bar{s}^{(3)}$	$p^{(4)}$	$\bar{s}^{(4)}$	$p^{(5)}$	$\bar{s}^{(5)}$	$p^{(6)}$	$\bar{s}^{(6)}$	$p^{(7)}$	$\bar{s}^{(7)}$
227	0.0944	1580	0.0108	6831	0.0022	11 949	0.0012	16 757	0.0008	19 249	0.0007

3.6. Characteristics of Hessian matrices

At the highest granularity level ($l = 7$), the design matrix $X^{(7)}$ is sparse and binary. Efficient computation of the Hessian matrix $X^{(7)\prime}X^{(7)}$ is enabled by NumPy's sparse matrix operations [47]. As noted in Remark 2, the diagonal entries of this Hessian matrix correspond to the frequencies of individual codes in $\alpha^{(7)}$. The panel **F** of Fig. 4 shows a log–log histogram of these diagonal entries, which follows a power-law distribution with exponent $\alpha \approx 1.93$ [48]. We note that the power-law distributions also appear in the subsets for FY-2019 and 2020. This behavior, reminiscent of word frequency distributions in natural language [48], suggests strong inter-code correlations [49].

We construct design matrices $X^{(l)}$ for $l = 2, 3, \ldots, 7$ using a custom Python function. The corresponding log–log histograms of the Hessian diagonals are also shown in Fig. 4. As granularity decreases (i.e., l becomes smaller), the number of columns $p^{(l)}$ is reduced due to code merging. While computing the full eigenvalue spectrum of the Hessian (needed for the effective dimension ρ in Eq. (9) for Ridge regression) is computationally intensive, its trace – equal to the sum of the diagonal entries – can be efficiently computed. Table 1 lists the values of $p^{(l)}$ and traces of the Hessian matrices for the FY2018 dataset. Consistent with Lemma 2, the trace increases as l decreases. The mean eigenvalue, $\bar{s}^{(l)} := \frac{\mathrm{tr}(X^{(l)\prime}X^{(l)})}{np^{(l)}}$, also captures this trend, which is included in Table 1 and used to estimate the upper bound ρ_B (Eq. (10)), as discussed in Lemma 1.

3.7. HCC and DRG code groupings

Hierarchical Condition Categories (HCC) offers a standardized method for grouping ICD-10 codes and is widely used in risk adjustment models. A many-to-one mapping exists between ICD-10-CM codes and HCC codes, as documented by the Centers for Medicare & Medicaid Services (CMS) [50]. There are 75 unique HCC codes, but only 4015 ICD-10-CM codes are included in these HCCs. The others are discarded. The left panel in Fig. 5 illustrates the frequency distribution of HCC codes appeared in our subset. In contrast to Panel F of Fig. 4, the HCC-based representation attenuates the power-law distribution observed in ICD-10 frequencies, where rare codes vastly outnumber common ones.

An alternative approach to aggregating diagnostic information is through Diagnosis-Related Groups (DRGs), which map principal diagnosis codes to a single DRG code. The DRG code for inpatient stay is available within the MedPAR dataset. From the subset analyzed, we identify 744 unique DRG codes. The right panel of Fig. 5 depicts the frequency distribution of these DRG codes. Similarly to the HCC codes, the long-tail pattern characteristic of ICD-10 code frequencies is notably diminished, indicating a reduction in the prevalence of rare codes under the DRG scheme.

4. Results

This section presents the results of regression analyses using the FY2018 dataset. We evaluate the log-linear model's predictive accuracy and coefficient stability across varying diagnostic code granularities and regularization strengths. Predictive performance is assessed under training–testing splits ranging from 1% to 60%, while coefficient stability is evaluated using repeated 80–20 subsampling without replacement to ensure distinct data partitions. For comparison, we also report results from DRG and HCC groupings, as well as decision tree and random forest models. All analyses were performed using Scikit-Learn [51].

4.1. Model accuracy and effective dimension

With the extracted design matrices $X^{(l)}$, it is straightforward to apply LinearRegression (OLS) and Ridge to estimate model coefficients and compute training and test R^2 scores. We vary the training set size by sampling training ratios $[0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6]$ from the full FY2018 dataset of $n = 495,023$ hospital stays. We first focus on the highest granularity case, $l = 7$. Fig. 6 summarizes the

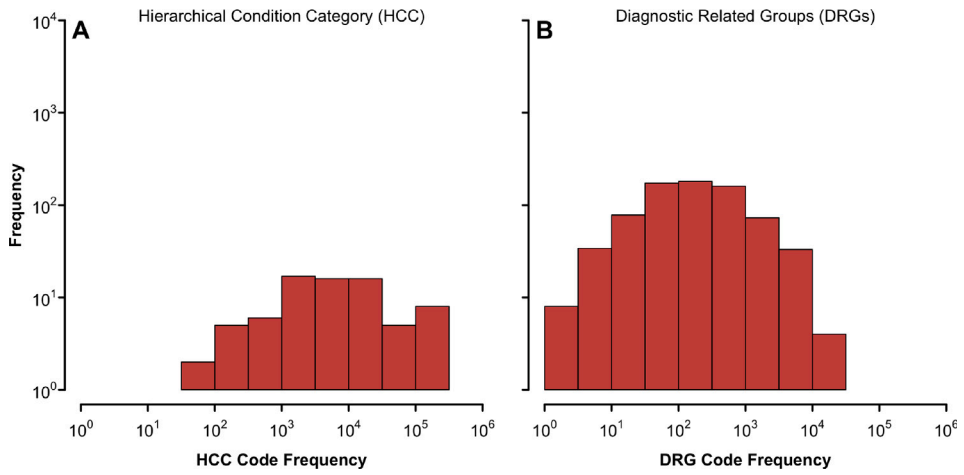


Fig. 5. Left panel: Histogram of HCC codes frequencies. Right panel: Histogram of DRG codes frequencies. Code groupings reduce the prevalence of rare codes in the ICD-10 representation of diagnoses.

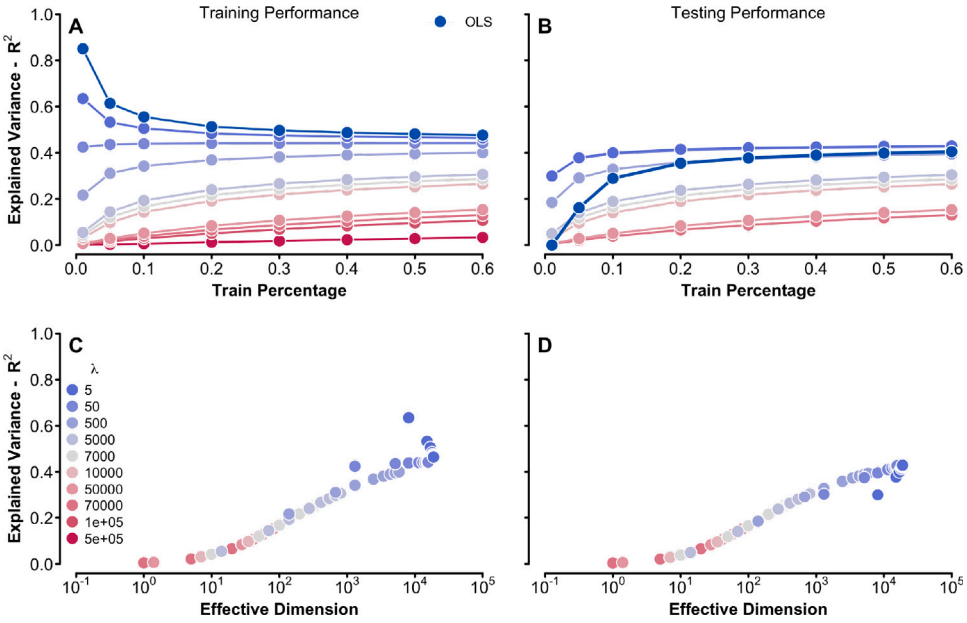


Fig. 6. Training (**A**) and test (**B**) scores for Ridge models with varying training size ratio and regularization λ. Panels (**C**) and (**D**) display these training and test, respectively, scores against the effective dimension upper bound ρ_B (Eq. (10)).

results for Ridge models trained with regularization strengths $\lambda = \{5, 50, 500, 5E3, 7E3, 1E4, 5E4, 7E4, 1E5, 5E5\}$. Training scores are shown in panel A, with OLS results in brightest blue for reference, while panel B shows the corresponding test scores. Overfitting is observed only for the smallest training size (approximately 5000) when $n_{tr} < p$ and regularization is weakest — evidenced by a large discrepancy between training and test scores. As expected, increasing regularization suppresses both scores, as the estimated $\hat{\beta}$ is increasingly biased toward zero [42]. Notably, both training and test scores decrease with smaller training ratios. This behavior arises because Scikit-Learn's Ridge implementation minimizes the unscaled objective, $\mathcal{L} = \sum(y_i - \hat{y}_i)^2 + \lambda\|\beta\|^2$, making the regularization term effectively stronger as the training set shrinks. However, when scores are plotted against the effective dimension upper bound $\rho_B = \frac{(p+1)\bar{s}}{\bar{s}+\lambda/n_{tr}}$, the results largely collapse onto a single curve, as shown in panels C and D of Fig. 6. Deviations from this collapse suggest that the sample covariance $X'X/n$ may be a noisy estimator of the true covariance for small training sizes.

Next, we examine how predictive performance varies with ICD-10 code granularity. Fig. 7 presents the training and test R^2 scores for Ridge regression models fitted to $X^{(l)}$ with truncation levels $CL \leq 2$ to $CL \leq 7$, using a fixed training ratio of 0.8. Ridge models were trained across eight regularization strengths within the range $\log_{10}\lambda \in [3.5, 6]$. Across all levels of granularity, OLS models achieve higher test scores than Ridge models. As ICD-10 codes are progressively aggregated to less granular representations – and as the regularization parameter increases – predictive accuracy decreases accordingly. For comparison, the model based on HCC groupings yields both training and test R^2 scores near 0.075, whereas the DRG-based model attains a training score of 0.41 and a test score of approximately 0.40, comparable to the OLS model using the most granular codes. The corresponding root-mean-squared error (RMSE) can be derived from the relationship $R^2 = 1 - \text{RMSE}^2/\text{TotalVariance}$ [5], with the total variance of the log-transformed cost equal to 0.13.

For completeness, we also report the predictive performance of tree-based models under the highest code granularity level ($l = 7$). A regularized decision tree model with min_sample_split = 100 achieved training and test R^2 scores of 0.38 and 0.09, respectively. A Random Forest model with n_estimators = 20 and the same regularization

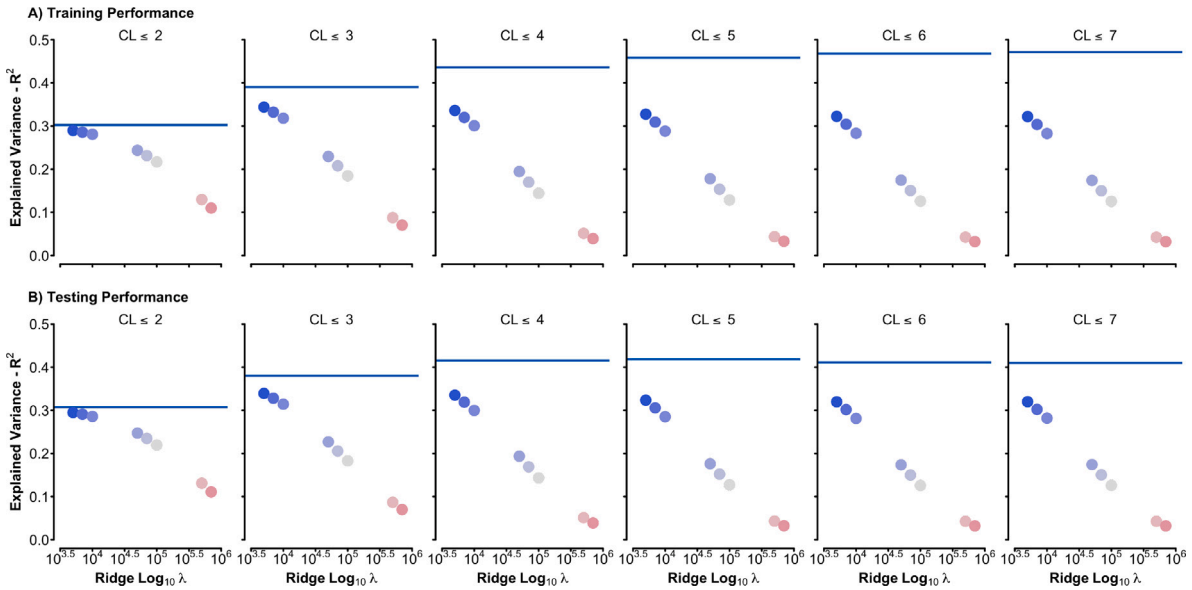


Fig. 7. Training (top) and test (bottom) predictive performance R^2 scores against regularization strength ($\log_{10}\lambda$) for log-linear models with different granularity levels specified by $CL \leq l = [2, 3, 4, 5, 6, 7]$. The bars denote the OLS result for the associated situation.

setting yielded corresponding scores of 0.46 (training) and 0.29 (test). These results suggest that while ensemble methods such as Random Forest improve predictive stability compared to a single decision tree, their performance remains comparable to the regularized linear models when applied to highly granular ICD-derived variables.

4.2. Coefficient consistency across samples and truncation levels

We assess the consistency of coefficient estimates across different training subsets by computing the consistency metric η (Eq. (11)) using an ensemble of 10 OLS models fit to randomly drawn training sets from the FY2018 data. The levels of truncation are also varied to show improved η (Eq. (11)) as codes become less granular. A consistent model should yield similar $\hat{\beta}$ vectors across different samples. This can be visualized with a scatterplot where one coefficient vector is drawn against the other. Perfect agreement corresponds to all points falling on the diagonal. Fig. 8 shows scatter plots for two sample pairs of coefficient vectors across all granularity levels. At the finest granularity ($l = 7$, bottom right), the coefficient clouds form rotated ellipses aligned with the diagonal, indicating *weak* agreement between samples. As granularity decreases (moving from bottom right to top left), the number of predictors $p^{(l)}$ drops, and the coefficient magnitudes shrink, mimicking the behavior of explicit regularization despite no penalty being applied in OLS. Importantly, the coefficient clouds are better explained by the diagonal line at lower granularity, meaning there is *strong* agreement between samples. To quantify this behavior, we compute η defined in Eq. (11) using the Spearman correlation from the SciPy library [52]. Fig. 9 displays the results. The left panel shows an increasing trend of η defined in Eq. (11) as the granularity is reduced in OLS models, suggesting more stable coefficients in lower-dimensional settings. The right panel displays consistency for Ridge models at $l = 7$, across eight regularization strengths $\lambda/n_{\text{tr}} = \{0, 2\text{E}{-6}, 2.5\text{E}{-6}, 2.5\text{E}{-5}, 1.3\text{E}{-4}, 2.5\text{E}{-4}, 1.3\text{E}{-3}, 2.5\text{E}{-3}\}$. Even small regularization leads to a noticeable jump (0.05) in η over OLS. As λ increases, consistency improves, reaching values near 0.9 for the strongest penalties. Since many ICD codes are rare, the diagonal entries of the Hessian – approximating eigenvalues – vary significantly. In the Ridge solution, $\left(\tilde{X}'\tilde{X}/n_{\text{tr}} + \lambda/n_{\text{tr}}\right)^{-1}$, the penalty term has minimal effect on coefficients associated with frequent codes (e.g., appearing 10,000+ times), but strongly dampens coefficients for infrequent ones. This selective shrinkage reduces coefficient variance and improves inter-sample agreement, hence increasing η.

As a comparison, the HCC and DRG code grouping schemes are also employed to fit the log cost with the method of OLS. The number of predictors is reduced from 19 249 (the finest granularity level using ICD-10 code) to 75 (HCC) and 744 (DRG), respectively. We shall emphasize that the design matrix X_{HCC} and X_{DRG} remain binary while the X's for lower granularity levels ($l < 7$) can have entries larger than unity in the process of variable aggregation. Unlike HCC and DRG, our scheme of code grouping does not discard any ICD code. The left panel in Fig. 10 displays the quantitative agreement between regression coefficients from training with different subsamples. The middle panel in Fig. 10 shows the result for DRG grouping. As both established schemes suppress the abundance of rare codes, the obtained consistency η in the right panel demonstrates improvement from directly employing the ICD-10 codes as predictors.

This analysis serves as an ablation study examining how varying ICD-10 code truncation levels affect predictive accuracy, coefficient stability, and interpretability. As shown in Fig. 7, OLS models reach their highest test accuracy (0.41) with fully detailed ICD-10 codes, declining gradually to about 0.30 when code length is reduced to two characters or fewer. Fig. 8 illustrates that lower granularity improves coefficient consistency, defined in Eq. (11), across subsamples, as the coefficient scatter tightens from a broad cloud to points concentrated along the diagonal. Similarly, Fig. 9A shows the stability metric increasing from roughly 0.75 to 0.9 with progressive truncation. Ridge regression and clinically grouped schemes such as DRG and HCC, however, maintain consistently higher stability across all levels.

5. Discussion

This study investigated how varying ICD code granularity and model regularization affected prediction accuracy and coefficient stability in log-linear models fitted to healthcare (inpatient) data. Our analysis of the Hessian matrix $X^{(7)\prime}X^{(7)}$ revealed that its diagonal entries follow a power-law distribution ($\alpha \approx 1.93$), reflecting a highly skewed code frequency landscape. This imbalance causes instability in coefficients associated with rare codes, which can be mitigated through implicit regularization via code aggregation or explicit regularization using Ridge regression. We also introduced the effective dimension upper bound ρ_B (Eq. (10)), which unifies model accuracy across varying

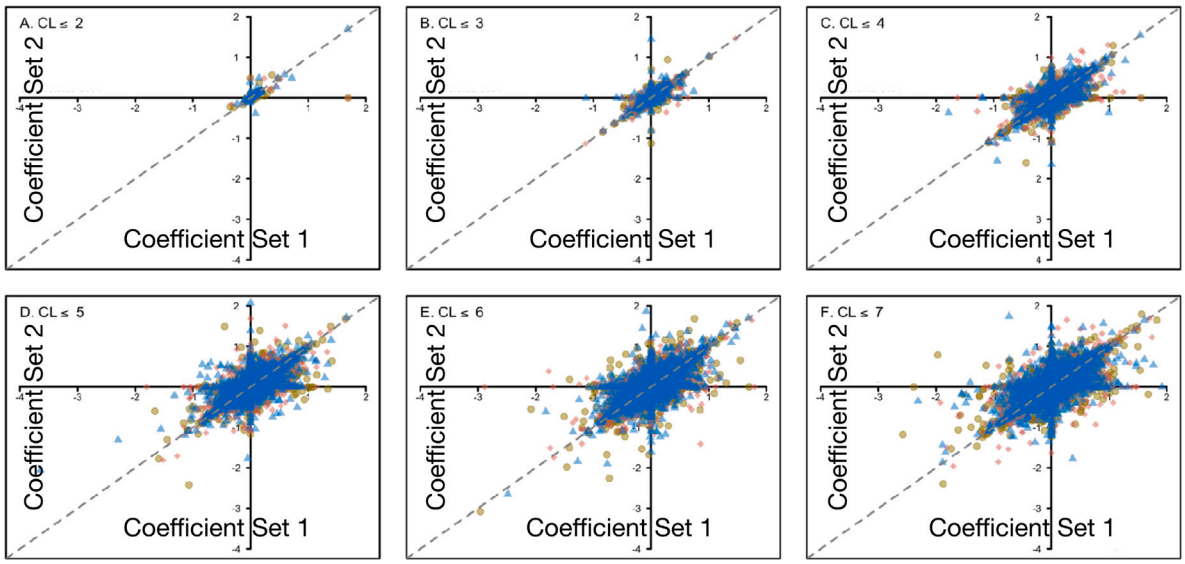


Fig. 8. Scatter plots comparing pairs of regression coefficients across resampled datasets, illustrating how code truncation improves coefficient stability. Panels correspond to truncation levels $CL \leq l = [2, 3, 4, 5, 6, 7]$, arranged from left to right, top to bottom. See the first paragraph of Section 4.2 for detailed discussion.

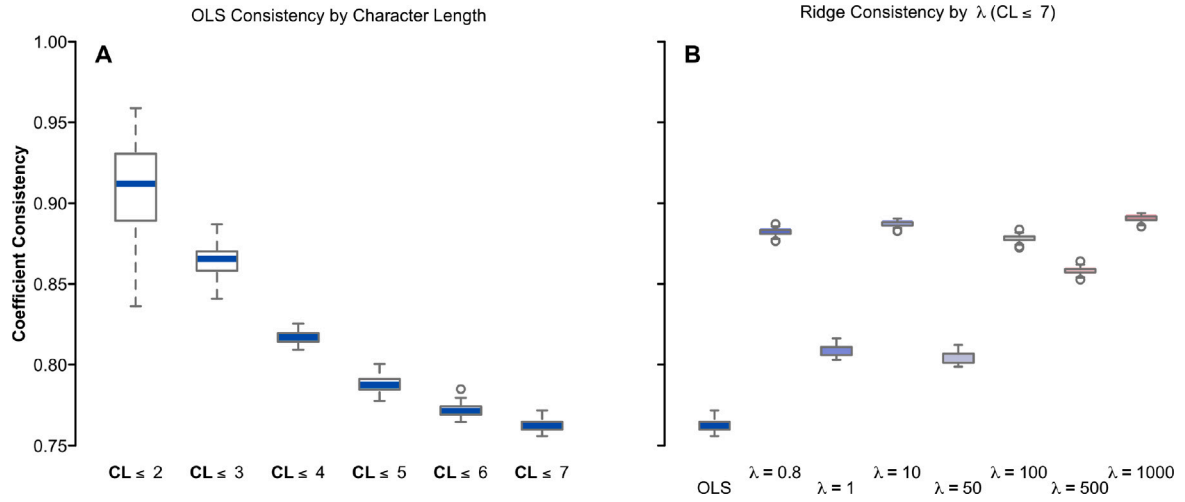


Fig. 9. Panel A: Coefficient consistency (Spearman correlation η, defined in Eq. (11)) for Ordinary Least Squares (OLS) models across varying ICD-10 code granularities ($CL \leq [2, 3, 4, 5, 6, 7]$) using a training ratio of 0.8. Increasing code truncation (i.e., lower granularity) improves coefficient stability, as reflected by higher η values. Panel B: Coefficient consistency for Ridge regression models with varying regularization strengths ($\log_{10} \lambda \in [3.5, 6]$) at the highest code granularity level ($CL \leq 7$). Stronger regularization leads to smoother and more stable coefficient estimates across subsampled datasets.

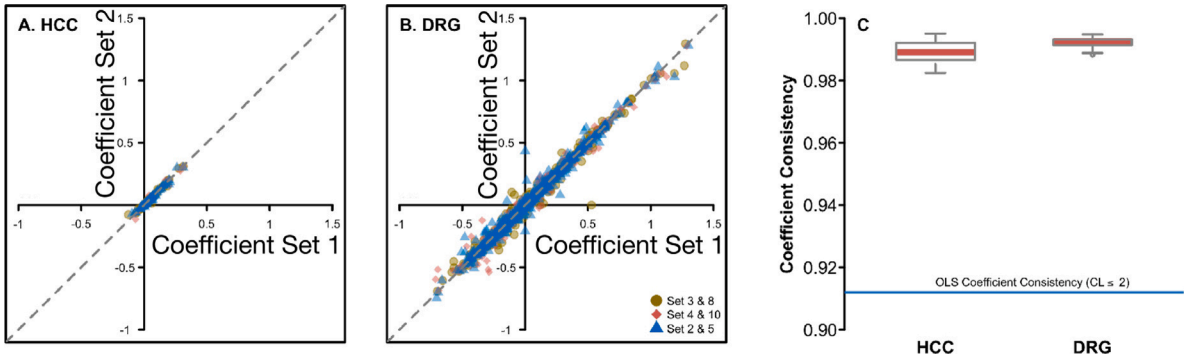


Fig. 10. Left panel: similar to Fig. 8, one set of the regression coefficients using HCC code groupings is scatter-plotted against another. Quantitative agreement between coefficients from training OLS with different subsamples is shown if the point clouds are concentrated to the diagonal line. Middle panel: result from employing DRG scheme. Right panel: the coefficient consistency η for both established schemes.

sample sizes and regularization levels. Plotting R^2 scores against the effective dimension ρ_B led to an empirical collapse of performance curves, suggesting it as a meaningful proxy for model capacity in sparse settings. Lastly, the proposed truncation framework can serve as a preprocessing layer for models such as random forests, gradient boosting, or neural networks by reducing dimensionality and providing more stable, interpretable inputs.

For a more practical implementation of code grouping aimed at reducing η, defined in Eq. (11), several important questions remain for future investigation. Given the critical importance of both accuracy and consistency in deploying machine learning models within the healthcare domain, future work should focus on strategies to balance these two aspects effectively. This study intentionally omits demographic variables in the section of Methods, as its primary goal is to examine how high-dimensional, sparse binary variables influence the consistency of regression coefficients. However, conducting subgroup analyses, e.g. races and specialties, to evaluate model accuracy and consistency remains particularly relevant for real-world healthcare applications. On the theoretical front, a promising direction involves modeling the underlying distribution from which highly correlated binary ICD-10 codes for inpatients are drawn. Deriving accuracy bounds [42,43,53] for models trained on such distributions would also be a valuable contribution.

Limitations. Our approach merges ICD codes based solely on code granularity (truncation or aggregation), which is just one strategy to reduce dimensionality. While systematic and reproducible, it does not use clinical knowledge to identify functionally or causally related codes. In contrast, real-world systems like CMS-HCC rely on expert-curated groupings to ensure interpretability and policy relevance, which our method does not attempt to replicate.

The analysis uses a geographically restricted subset of MedPAR FY2018, covering only Downstate New York providers. Hospitals in this region are larger, costs are higher, and patient demographics differ from national averages. Therefore, OLS coefficients estimated from this subset may not generalize to other regions. While our method is data-agnostic, validation on additional datasets or years is necessary.

The space of possible ICD code groupings is combinatorially large, and grouping choices can greatly affect model performance, interpretability, and fairness. We treat granularity as a proxy for grouping, but discovering optimal, data-driven groupings that stabilize coefficients and maintain predictive accuracy remains an open challenge. Future work could explore automated grouping strategies, similar to learned pooling in computer vision, to produce more robust and adaptive models [46].

Other limitations include focusing on a single year of data and excluding non-diagnostic predictors such as procedures or medications from the main stability analysis. While we report predictive performance for tree-based models like decision trees and random forests, we cannot make definitive claims about coefficient stability in these models. Unlike OLS, these models lack a clear theoretical framework, such as Lemma 2. Extending stability analysis to complex machine learning models is an important direction for future research.

6. Conclusions

In this study, we employed the OLS method to model the logarithm of hospital stay costs using sparse indicators derived from ICD-10 codes. We identified a key issue: the instability of regression coefficients across different training subsamples. To address this, we proposed a metric based on Spearman correlation to quantify coefficient inconsistency. Since OLS coefficients can serve as quantifications of individual patient health risks, ensuring consistency is as crucial as achieving accuracy and interpretability in healthcare-related machine learning models. To mitigate inconsistency, we introduced a truncation approach that groups ICD-10 codes, thereby aggregating regression variables. From a mathematical standpoint, small eigenvalues in the Hessian matrix is due to infrequent codes, contributing to high variance in OLS coefficient estimation. By reducing the granularity of ICD-10 codes through grouping, we effectively increase the trace of the Hessian matrix, acting as an implicit form of regularization that enhances coefficient stability. Furthermore, by examining the distribution of diagonal entries in the Hessian matrices, we observed that reducing code granularity alleviates the overrepresentation of rare codes. Established grouping schemes such as Hierarchical Condition Categories (HCC) and Diagnosis-Related Groups (DRG) similarly help mitigate the sparsity caused by infrequent codes, contributing to robust, interpretable, and practical risk adjustment models. In summary, the proposed truncation-based regularization provides a conceptual bridge between interpretable OLS models and modern machine learning methods, enabling stability and dimensionality reduction while maintaining transparency of diagnostic features.

CRediT authorship contribution statement

Chi-Ken Lu: Conceptualization of this study. Development of methodology, theory, proofs, and software. Manuscript drafting and editing. **David Alonge:** Preliminary accuracy study of OLS. **Nicole Richardson:** Compilation of cost charge ratio for providers in NYC. **Bruno Richard:** Manuscript drafting and editing. Cost charge ratio compilation. Conceptualization of the study and methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments and funding sources

The encouragements and suggestions from anonymous reviewers of MLHC 2025 are greatly appreciated. This work was supported by The Analytics for Equity Initiative under National Science Foundation (award number: 49100423C0041).

Data availability

The data used in this study are from the Medicare Provider Analysis and Review (MedPAR) file, which contains administrative claims data for services provided to Medicare beneficiaries. These data are not publicly available due to patient privacy protections but can be obtained through a data use agreement with the Centers for Medicare & Medicaid Services (CMS). Interested researchers may request access by submitting a research protocol and data request through the CMS Research Data Assistance Center (ResDAC).

References

[1] Naihua Duan, Willard G. Manning, Carl N. Morris, Joseph P. Newhouse and, A comparison of alternative models for the demand for medical care, J. Bus. Econom. Statist. 1 (2) (1983) 115–126.
[2] Michael Griswold, Giovanni Parmigiani, Arnie Potosky, Joseph Lipscomb, Analyzing health care costs: a comparison of statistical methods motivated by medicare colorectal cancer charges, Biostatistics 1 (1) (2004) 1–23.
[3] Timothy J. Layton, Imperfect risk adjustment, risk preferences, and sorting in competitive health insurance markets, J. Health Econ. 56 (2017) 259–280.
[4] Iván Sánchez Fernández, Marta Amengual-Gual, Cristina Barcia Aguilar, Tobias Loddenkemper, Estimating the cost of status epilepticus admissions in the United States of America using ICD-10 codes, Seizure 71 (2019) 295–303.
[5] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning, Springer Texts in Statistics, Springer New York, NY, 2021.
[6] Hong J Kan, Hadi Kharrazi, Hsien-Yen Chang, Dave Bodycombe, Klaus Lemke, Jonathan P Weiner, Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults, PLoS One 14 (3) (2019) e0213258.

[7] Jeremy A Irvin, Andrew A Kondrich, Michael Ko, Pranav Rajpurkar, Behzad Haghgoo, Bruce E Landon, Robert L Phillips, Stephen Petterson, Andrew Y Ng, Sanjay Basu, Incorporating machine learning and social determinants of health indicators into prospective risk adjustment for health plan payments, BMC Public Health 20 (2020) 1–10.

[8] S.D. Reed, D.K. Blough, K. Meyer, J.G. Jarvik, Inpatient costs, length of stay, and mortality for cerebrovascular events in community hospitals, Neurology 57 (2) (2001 Jul 24) 305–314.

[9] Wichayaporn Thongpeth, Apiradee Lim, Akemat Wongpairin, Thaworn Thongpeth, Santhana Chaimontree, Comparison of linear, penalized linear and machine learning models predicting hospital visit costs from chronic disease in Thailand, Inform. Med. Unlocked 26 (2021) 100769.

[10] J Ruth Sandra, Sanjana Joshi, Aditi Ravi, Ashwini Kodipalli, Trupthi Rao, Shoaib Kamal, Prediction of cost for medical care insurance by using regression models, in: International Conference on Emerging Research in Computing, Information, Communication and Applications, Springer, 2023, pp. 311–323.

[11] A. Ravishankar Rao, Raunak Jain, Mrityunjai Singh, Rahul Garg, Predictive interpretable analytics models for forecasting healthcare costs using open healthcare data, Healthc. Anal. 6 (2024) 100351.

[12] Gregory C. Pope, John Kautter, Melvin J. Ingber, Sara Freeman, Rishi Sekar, Cordon Newhart, Melissa A. Evans, Evaluation of the CMS-HCC Risk Adjustment Model, Technical Report, The Centers for Medicare and Medicaid Services Office of Research, Development, and Information, 2011.

[13] John Kautter, Gregory C Pope, Melvin Ingber, Sara Freeman, Lindsey Patterson, Michael Cohen, Patricia Keenan, The HHS-HCC risk adjustment model for individual and small group markets under the Affordable Care Act, Medicare Medicaid Res. Rev. 4 (3) (2014) mmrr2014–004.

[14] Arlene S. Ash, Eric O. Mick, Randall P. Ellis, Catarina I. Kiefe, Jeroan J. Allison, Melissa A. Clark, Social determinants of health in managed care payment formulas, JAMA Intern. Med. 177 (10) (2017) 1424–1430.

[15] Arthur E. Hoerl, Robert W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67.

[16] Trevor Hastie, Robert Tibshirani, Jerome Friedman, The elements of statistical learning: Data mining, inference, and prediction, 2017.

[17] Ian Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning, MIT Press, 2016, http://www.deeplearningbook.org.

[18] A. Pfuntner, L. Wier, C. Steiner, Costs for hospital stays in the United States, in: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs, Agency for Healthcare Research and Quality, Rockville (MD; US), 2006-02, pp. 1–11, Number 146.

[19] Centers for Medicare & Medicaid Services, ICD-10 files & news archive, 2025, https://www.cms.gov/medicare/coding-billing/icd-10-codes/icd-10-cm-icd-10-pcs-gem-archive. (Accessed 23 July 2025).

[20] Kimberly J. O'Malley, Karon F. Cook, Matt D. Price, Kimberly Raiford Wildes, John F. Hurdle, Carol M. Ashton, Measuring diagnoses: ICD code accuracy, Health Serv. Res. 40 (5p2) (2005) 1620–1639.

[21] Nisreen Shiban, Joshua Gaul, Henry Zhan, Andrew Elhabr, Nima Kokabi, Jamlik-Omari Johnson, Tarek Hanna, Justin Schrager, Judy Gichoya, Imon Banerjee, Hari Trivedi, Machine learning methods to predict survival in patients following traumatic aortic injury, MedRxiv (2021).

[22] Mark Mirtchouk, Bharat Srikishan, Samantha Kleinberg, Hierarchical information criterion for variable abstraction, in: Machine Learning for Healthcare Conference, PMLR, 2021, pp. 440–460.

[23] Edmund M. Qiao, Alexander S. Qian, Vinit Nalawade, Rohith S. Voora, Nikhil V. Kotha, Lucas K. Vitzthum, James D. Murphy, Evaluating high-dimensional machine learning models to predict hospital mortality among older patients with cancer, JCO Clin. Cancer Inform. (6) (2022) e2100186, PMID: 35671416.

[24] Dimitrios Zikos, Nailya DeLellis, Comparison of the predictive performance of medical coding diagnosis classification systems, Technologies 10 (6) (2022).

[25] Robert B. Fetter, Jean L. Freeman, Diagnosis related groups: Product line management within hospitals, Acad. Manag. Rev. 11 (1) (1986) 41–54.

[26] William J. Lynk, One DRG, one price? The effect of patient condition on price variation within DRGs and across hospitals, Int. J. Health Care Finance Econ. 1 (2) (2001) 111–137.

[27] Todd H. Wagner, Anjali Upadhyay, Elizabeth Cowgill, Theodore Stefos, Eileen Moran, Steven M. Asch, Peter Almenoff, Risk adjustment tools for learning health systems: A comparison of DxCG and CMS-HCC V21, Health Serv. Res. 51 (5) (2016) 2002–2019.

[28] Juyoung Kim, Minsu Ock, In-Hwan Oh, Min-Woo Jo, Yoon Kim, Moo-Song Lee, Sang-il Lee, Comparison of diagnosis-based risk adjustment methods for episode-based costs to apply in efficiency measurement, BMC Health Serv. Res. 23 (1) (2023) 1334.

[29] Joseph C. Gardiner, Zhehui Luo, Cathy J. Bradley, Corina M. Sirbu, Charles W. Given, A dynamic model for estimating changes in health status and costs, Stat. Med. 25 (21) (2006) 3648–3667.

[30] David A. Jenkins, Matthew Sperrin, Glen P. Martin, Niels Peek, Dynamic models to predict health outcomes: current status and methodological challenges, Diagn. Progn. Res. 2 (1) (2018/12/18) 23.

[31] Willard G. Manning, John Mullahy, Estimating log models: To transform or not to transform? J. Health Econ. 20 (4) (2001) 461–494.

[32] Robert Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (2018) 267–288.

[33] Sergey Bakin, Adaptive Regression and Model Selection in Data Mining Problems (Ph.D. thesis), The Australian National University, 1999.

[34] Ming Yuan, Yi Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc. Ser. B Stat. Methodol. 68 (1) (2006) 49–67.

[35] Ahmed I. Taloba, Rasha M. Abd El-Aziz, Huda M. Alshanbari, Abdal-Aziz H. El-Bagoury, Estimation and prediction of hospitalization and medical care costs using regression in machine learning, J. healthc. Eng. 2022 (1) (2022) 7969220.

[36] Benedikt Langenberger, Timo Schulte, Oliver Groene, The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data, PLoS One 18 (1) (2023) 1–16.

[37] Chengliang Yang, Chris Delcher, Elizabeth Shenkman, Sanjay Ranka, Machine learning approaches for predicting high cost high need patient expenditures in health care, BioMed. Eng. OnLine 17 (1) (2018) 131.

[38] Yead Rahman, Prerna Dua, A machine learning framework for predicting healthcare utilization and risk factors, Healthc. Anal. 8 (2025) 100411.

[39] Alnur Ali, Edgar Dobriban, Ryan Tibshirani, The implicit regularization of stochastic gradient flow for least squares, in: Hal Daumé III, Aarti Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 233–244.

[40] Gauthier Gidel, Francis Bach, Simon Lacoste-Julien, Implicit regularization of discrete gradient dynamics in linear neural networks, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc., 2019.

[41] Lee H. Dicker, Variance estimation in high-dimensional linear models, Biometrika (2014) 269–284.

[42] Gabriel Mel, Surya Ganguli, A theory of high dimensional regression with arbitrary correlations between input features and target functions: sample complexity, multiple descent curves and a hierarchy of phase transitions, in: International Conference on Machine Learning, PMLR, 2021, pp. 7578–7587.

[43] Trevor Hastie, Andrea Montanari, Saharon Rosset, Ryan J Tibshirani, Surprises in high-dimensional ridgeless least squares interpolation, Ann. Stat. 50 (2) (2022) 949.

[44] Wesley J. Maddox, Gregory Benton, Andrew Gordon Wilson, Rethinking parameter counting in deep models: Effective dimensionality revisited, 2020, arXiv preprint arXiv:2003.02139.

[45] Shubhra Aich, Ian Stavness, Global sum pooling: A generalization trick for object counting with small datasets of large images, 2018, arXiv preprint arXiv:1805.11123.

[46] Manli Sun, Zhanjie Song, Xiaoheng Jiang, Jing Pan, Yanwei Pang, Learning pooling for convolutional neural network, Neurocomputing 224 (2017) 96–104.

[47] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, Travis E. Oliphant, Array programming with NumPy, Nature 585 (7825) (2020) 357–362.

[48] Aaron Clauset, Cosma Rohilla Shalizi, Mark E.J. Newman, Power-law distributions in empirical data, SIAM Rev. 51 (4) (2009) 661–703.

[49] Michael Mahoney, Charles Martin, Traditional and heavy tailed self regularization in neural network models, in: International Conference on Machine Learning, PMLR, 2019, pp. 4284–4293.

[50] Centers for Medicare & Medicaid Services, 2018 model software ICD-10 mappings, 2025, https://www.cms.gov/medicare/health-plans/medicareadvtgspecratestats/risk-adjustors-items/risk2018. (Accessed 23 July 2025).

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[52] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R.J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental algorithms for scientific computing in python, Nature Methods 17 (2020) 261–272.

[53] Zhenyu Liao, Romain Couillet, Michael W. Mahoney, A random matrix analysis of random fourier features: Beyond the gaussian kernel, a precise phase transition, and the corresponding double descent, Adv. Neural Inf. Process. Syst. 33 (2020) 13939–13950.