Cluster Analysis reveals Socioeconomic Disparities among Elective Spine Surgery Patients

Alena Orlenko*[1], Philip J. Freda*[1], Attri Ghosh[1], Hyunjun Choi[1], Nicholas Matsumoto[1], Tiffani J. Bright[1], Corey T. Walker[2], Tayo Obafemi-Ajayi[3], Jason H. Moore[1]

[1]*Department of Computational Biomedicine*
[2]*Department of Neurosurgery*
Cedars-Sinai Medical Center, Los Angeles, California, USA

[3]*Engineering Program*
Missouri State University, Springfield, Missouri, USA

This work demonstrates the use of cluster analysis in detecting fair and unbiased novel discoveries. Given a sample population of elective spinal fusion patients, we identify two overarching subgroups driven by insurance type. The Medicare group, associated with lower socioeconomic status, exhibited an over-representation of negative risk factors. The findings provide a compelling depiction of the interwoven socioeconomic and racial disparities present within the healthcare system, highlighting their consequential effects on health inequalities. The results are intended to guide design of fair and precise machine learning models based on intentional integration of population stratification.

Keywords: clustering; fairness; equity; explainability; feature importance; informatics.

1. Introduction

Advances in machine learning (ML) technologies paralleled with increased clinically relevant data availability have led to major progress in precision medicine over the past decade.[1] Data-driven solutions, particularly ML methods, are becoming integral to personalized predictive medicine as they can inform clinical decision support systems, generate accurate patient risk stratification models, and contribute to intelligent guideline development using high-dimensional complex medical data.[2] Indeed, ML-based approaches have generated robust predictive models in the diagnoses of several diseases such as cardiovascular diseases,[3] type II diabetes,[4] and early-stage Alzheimer's disease[5] and for post-surgical outcomes and treatment response in several procedures including cardiac surgery[6] and spinal surgeries.[2,7,8] Thus, clinicians can utilize this information to evaluate risk of poor diagnoses and adverse outcomes, assisting clinical decision making by providing personalized assessments of the benefits and consequences related to undergoing or delaying invasive procedures.

The rates of spine surgery, an invasive procedure, have been steadily increasing over the past few decades.[9] With the proportion of the elderly population projected to dramatically

*These authors contributed equally to the paper.

This work was supported by National Institutes of Health (USA) grant R01 LM010098.

increase in the coming years, utilization of spinal procedures is expected to follow as degenerative spine conditions become more prevalent.[10] Spinal fusions generally require extensive muscle dissection and reconstruction of the spinal column, which typically necessitates significant post-operative opioid consumption and comes with considerable post-operative risks.[11] With the potential for long recovery periods and the risk of the development of opioid dependency as a result of these surgeries, outcome prediction in spinal fusion surgeries has become an important area of research. To accurately predict outcomes, it is crucial to consider patient diversity, which stems from various sources, including but not limited to biological, societal, environmental, and psychosocial factors.[12] These sources of diversity can result in significantly different outcomes, ultimately affecting a patient's long-term quality of life after surgery.

For data-driven predictive models to become widely and safely adopted in clinical settings, key research challenges still remain to be resolved. These include assessing clinical heterogeneity and avoiding bias in decision-making. Complex ML algorithms have an inherent tendency for biased decisions that disproportionately impact underrepresented demographic groups leading to possible discriminatory outcomes.[13] This concern is frequently overlooked in study design, resulting in unequal treatment of minority individuals.[14] We seek to examine the intricate heterogeneity in clinical data to identify any differential patient subgroups, if present. This will enable us to mitigate bias in the ML decision-making for clinical systems.

Cluster analysis has been applied in a wide range of applications as an exploratory tool to enhance knowledge discovery.[15–17] It can help by identifying more homogeneous subgroups for effective ML models. The goal is to detect and characterize novel sub-types that exhibit differing clinical patterns and/or outcome trajectories that may benefit from different treatment options. Ultimately, the validity of any sub-grouping paradigm depends on whether the resulting sub-groups uncover/expose some biologic or genetic variation, which can be used to predict prognoses, recurrent risks, or treatment responses. However, most of the approaches employ a single clustering algorithm with limited explainability.[15–17] To overcome these limitations, we introduce a novel clustering framework to examine and characterize a cohort of patients that have undergone elective spinal fusion surgery at Cedars-Sinai Medical Center.

2. Data

The dataset consists of electronic health records (EHR) of 5,214 elective spinal fusion (ESF) surgery procedures derived from 4,930 patients (ages 18-85) at the authors' single institution from 2013 to 2022. Only patients who survived after surgery, with two or fewer procedures are included. If the second procedure was conducted within seven days of the first, the most recent is retained. Patients with a second procedure conducted after seven days but less than a year apart are excluded. Forty-five features from the patient's health records were selected and integrated in the cluster analysis. These features span baseline characteristics/demographics, pre-surgery clinical labs, vitals, medication lists, past medical history, post-operative care, and social status, as guided by domain expert (C.T.W.).

The race feature consolidates both self-reported race and ethnicity information. Self-reported ethnicity of "Hispanic", regardless of race, is represented as "Hispanic". Race designation of "Asian" or "Native Hawaiian or other Pacific Islander" are categorized as

"Asian/Pacific Islander". "Native American or Alaska Native", "Other", "Patient declined", "Unknown", or missing are all consolidated as "Other". Social status features include insurance type, marital status, smoking, and alcohol use. Patients with commercial or private insurance are grouped as "commercial" while Medicare, California's Medicaid program (Medi-Cal), or all other government insurance are categorized as "medicare". Vitals features include systolic blood pressure (SBP), body mass index (BMI) and pain score. We include the most frequently used lab value results from the EHR that had less than 50% of missing data (11: hemoglobin, white blood cell (WBC) count, red blood cell (RBC) count, platelet count, potassium, sodium, chloride, blood urea nitrogen (BUN), creatinine, calcium, and blood type) . Selected post-operative care features are discharge disposition, length of stay, and readmission status. Past medical history (PMH) features (yes/no) are derived by aggregating the ICD codes relevant to specific conditions of interest (metabolic, anxiety, chronic pain, mood, headache, nicotine, other psychiatric, opioid substance use disorder (SUD), alcohol SUD, cannabis SUD, and other SUD). Medication list features are derived based on usage of medications under 7 broad categories, as defined by the domain expert. These include muscle relaxers, non-opioid analgesic, psychiatric, sleep, medication-assisted treatment, gabapentinoids, and "other".

The summary of the baseline characteristics is presented in Table 1. For a complete list of the medications that map to each medication feature as well as the ICD codes that map to each PMH feature, see supplementary file [a]. Data request approved by the Cedars-Sinai Honest Enterprise Research Brokers (HERB) committee. This research study was carried out under the guidelines and approval of the Cedars-Sinai Institutional Review Board.

Table 1. Demographic summary of elective spinal fusion surgery patient sample (n = 5,214).

Characteristic	Distribution
Age	median: 67 range: 18 - 85; 65+: 57.59%
Gender	Male: 46.47% Female: 53.53%
Race	White: 75.66%, Hispanic: 10.32%, Black/African-American: 6.75%, Asian: 3.55%, Other: 3.72%
Insurance type	Medicare: 45.09% (65+: 88.74%) (Medicare: 96.85%, Medi-Cal: 0.02%, Other government: 0.01%) Commercial: 53.93% (65+: 31.80%), No Insurance: 0.98%
Marital status	Single: 17.97%, Married: 63.57%, Divorced: 9.51%, Widowed: 6.23%, Significant other: 2.51%, Unknown: 0.21%

3. Methods

To ensure a fair and unbiased model, we propose a robust automated system that integrates multiple clustering algorithms, ensemble internal validation metrics, automated ML (autoML)-driven explainability, and post-hoc univariate statistical analysis.

The data curation steps involve the detection of erroneous, non-biologically plausible values, and/or outliers. Domain expert guidance in conjunction with outlier analyses are applied to ensure mitigation of potential bias and possible human data entry errors. These values are dropped and imputed, rather than dropping the entire sample. Missing values are imputed using the multivariate feature imputation (*IterativeImputer* method in Python).[18] All 45 fea-

[a] Supplementary information is available at: https://github.com/EpistasisLab/PSB2024_spine/

tures are not highly intercorrelated as evident from passing the correlation filter analysis using the Pearson and Spearman rank correlations (≤ 0.85).

We perform an automated clustering method that incorporates hyperparameter sampling across various algorithms that permutes the distance type (Euclidean, Manhattan), and number of clusters (k=[2:10]), when applicable. It exploits five individual algorithms (Spectral, Agglomerative, k-means, Birch, and Gaussian mixture).[19] We also conduct an ensemble clustering model that leverages these individual methods using the mixture model consensus metric in *OpenEnsemble*.[20,21] Our model includes TooManyCells (TMC) spectral hierarchical clustering method,[22] for a total of seven methods with 68 permutations. To integrate TMC into the automated clustering pipeline, we implement an extension that aggregates cluster labels with multiple terminal cluster nodes starting at the root node. The depth of the tree partition serves as a TMC hyper-parameter. The optimal clustering output is determined using the ensemble internal validation metric model introduced by Nguyen et al..[23] The model assigns a final score based on a consensus of five metrics (Calinski-Harabasz, Davies-Bouldin, Silhouette score, $\mathcal{I}$, and Xie-Benie).[24] Each metric ranks its top 15 results and sets the remainder to zero. The ensemble model assigns a final overall rank score to each clustering outcome based on the weighted sum of the individual ranking assignment of each metric.

Key novelty of our clustering framework is that we utilize a model-agnostic approach to evaluate the feature importance and assess which key discriminant features are driving cluster separation with an autoML tool, TPOT.[25] TPOT evaluates the informative contributions of features to clustering results by predicting cluster labels with each feature independently. In contrast to the current state-of-the-art methods for evaluating feature importance (such as SHapley Additive exPlanation,[26] Permutation feature importance, Gini impurity in Random Forest[27]), TPOT overcomes the single model limitation as it searches and optimizes across multiple ML algorithms. For each feature, we run the TPOT optimization (across 13 different classifiers configuration), and extract the best-performing model performance as the feature importance metric. This provides insight into the key discriminant input features and guides the next steps of analysis. Visualization of results is performed using ISOMAP[28] and TMC dendograms. Code for all the methods are available at (https://github.com/EpistasisLab/PSB2024_spine).

Univariate global statistical tests are conducted, as post-hoc analyses, to assess which features exhibit differences among the cluster groups. The method of analysis differs depending on the measurement scale of the feature. Features with significant test results suggest utility in clustering. For continuous features, we test for normality using Shapiro-Wilk tests. All features are non-normally distributed. Thus, we employ non-parametric Mann-Whitney tests (or Kruskal-Wallis tests in case of multiple groups). For categorical and binomial features, we use Chi-square tests of independence. The resulting p-values of these tests are corrected for multiple testing using the Benjamini-Hochberg procedure.

4. Results

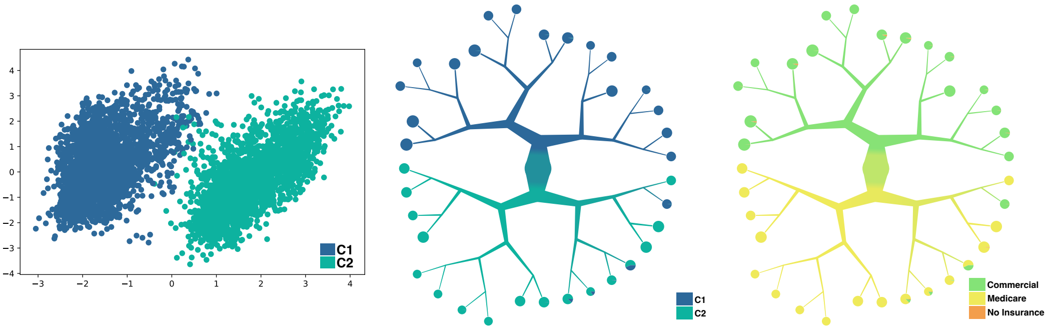
4.1. *Entire ESF sample is stratified by socioeconomic factor of insurance.*

Upon evaluating ensemble clustering on the overall cohort of 5,214 surgeries, Table 2A shows k-means with two clusters consistently outperforms other methods across internal validation

Table 2. Top ranked results for 1^{st} and 2^{nd} order clustering based ensemble validation rank scores.

Output [Cluster sizes]	**CH (rank)**	**Db (rank)**	**I (rank)**	**Sil (rank)**	**Xb (rank)**	**Overall Rank**
A. Clustering on entire cohort						
$k$means-2 [2852, 2362]	550.01 (15)	3.03 (14)	0.619 (15)	0.098 (15)	2.348 (15)	74
GaussianMixture-2 [2732, 2482]	549.56 (14)	3.03 (15)	0.619 (14)	0.097 (14)	2.348 (14)	71
Spectral (euclidean)-2 [2863, 2351]	533.62 (13)	3.08 (11)	0.597 (13)	0.095 (13)	2.436 (13)	63
B. 2^{nd} order clustering on C1 group						
$k$means-2 [1872, 980]	202.13 (15)	3.57 (0)	0.398 (15)	0.089 (14)	3.180 (3)	47
Spectral (manhattan)-2 [1931, 921]	168.17 (13)	3.86 (0)	0.341 (14)	0.081 (13)	3.705 (0)	40
Mixture model-2 [1638, 1214]	168.23 (14)	4.06 (0)	0.305 (13)	0.074 (12)	4.142 (0)	39
C. 2^{nd} order clustering on C2 group						
$k$means-2 [1476, 886]	204.92 (15)	3.27 (11)	0.503 (15)	0.094 (15)	2.700 (15)	69
GaussianMixture-2 [1474, 888]	204.88 (14)	3.27 (10)	0.503 (14)	0.094 (14)	2.702 (14)	64
Mixture model-2 [1473, 889]	195.48 (13)	3.36 (1)	0.480 (13)	0.094 (13)	2.834 (13)	54

metrics. Top ranking methods (k-means, Gaussian Mixture, spectral, TooManyCells) return similar 2-cluster partitions and display high consistency as top performers across all five metrics. Subsequent analyses are conducted on the k-means-2 result (C1 and C2). The visualization of the subgroups is shown using both ISOMAP (Figure 1(a)) and TMC (Figure 1(b)). Note: TMC performs its embedded technique (spectral hierarchical clustering) prior to visualization, hence, not representing C1 and C2 separation exactly. TPOT feature importance analysis reveals that insurance type, a potential socioeconomic factor, is most important to cluster separation explainability (100% balanced accuracy (B-Acc.)). Age, discharge disposition, and PMH metabolic are of less importance (79.1%, 64.2%, 62.7% B-Acc. respectively). Mapping the insurance type label with TMC dendrograms confirms this as well (Figure 1). Cluster C1 consists of all patients with "commercial insurance" and 40 of "no insurance" while C2 has all patients on "medicare insurance" and 11 with "no insurance".



(a) ISOMAP representation (b) C1 & C2 using TMC (c) TMC with insurance

Fig. 1. Visualization of k-means-2 results on entire cohort.

Age is a determinant for medicare eligibility (65+) in the USA. Thus, we conduct univariate statistical analyses between C1 and C2 (insurance-driven clusters) as well as between and within age-stratified subgroups. Figure 2 illustrates the experimental design of these analyses. The pairwise comparisons are conducted as follows: Exp 1: C1 vs. C2; Exp 2: 65+ subgroups of C1 and C2 i.e., C1 $\geq$ 65 vs. C2 $\geq$ 65; Exp 3: 65- subgroups of C1 and C2 i.e., C1 $<$ 65 vs. C2 $<$ 65; Exp 4: within C1: C1 $\geq$ 65 vs. C1 $<$ 65; Exp 5: within C2: C2 $\geq$ 65 vs. C2 $<$ 65.

4.1.1. *Univariate analysis reveals health disparities associated with insurance types.*

Figure 3 summarizes the key features that differ significantly at the entire cohort level between C1 and C2 and when age-stratified (Exp 1, 2, and 3). Nine features display age-independence as they are statistically different across all three comparisons (Figure 3a). These are race, marital status, discharge disposition, hemoglobin, platelet count, RBC count, potassium, and two PMH features (metabolic and anxiety). We also observe that there are some features that are not different between C1 and C2 (Exp 1), but do exhibit significant differences within the 65- comparisons (Exp 3) (Figure 3b). These features (PMH features of pain score, other psychiatric disorders, nicotine use, headache, other SUDs, and use of non-opioid analgesics) imply some possible health disparities between the two socioeconomic driven groups after accounting for the age factor. (Note, an additional significant feature, PMH of other SUD, isn't shown in the figure, as it affects less than 5% of the overall population.) There are no features that are significant only between 65+ subgroups (Exp 2) and not at the entire cohort level (Exp 1). See Supplementary file [b] for complete details of all the pairwise comparisons.

The analysis also reveals some features that are significant across all three comparisons (Exp 1, 2 and 3), which are also significant within C1 and C2 when stratified by age (Exp 4 and 5). These include race, platelet count, RBC count, marital status, discharge disposition, and PMH features of metabolic and anxiety (see Supplementary file[b]). Features such as hemoglobin are significant within C1 (Exp 4) but not C2 (Exp 5). All PMH features are significantly different within C2 age-stratified groups. Overall, negative health factors, such as lower hemoglobin, RBC, platelet count, potassium levels, and higher incidence of metabolic disease and anxiety are associated with C2, indicating socioeconomic health disparities.

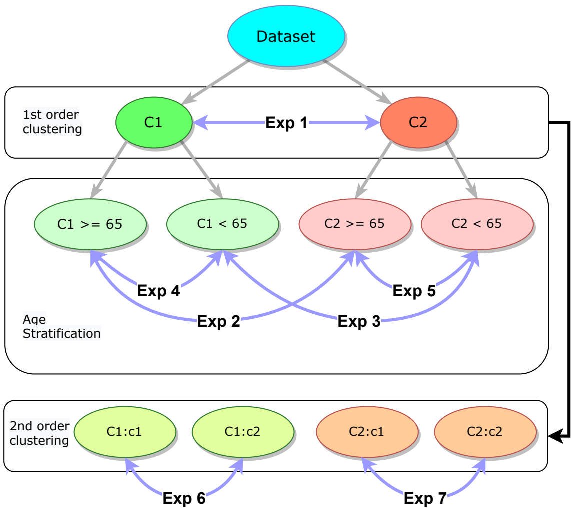


Fig. 2. Experimental design of cluster analyses and pairwise comparisons.

4.1.2. *Adverse outcomes are disproportionately observed in minority racial groups.*

From the pairwise comparisons (Exps 1-5), race is consistently significant. The 65- population in C2 had a larger proportion of non-white patients (60% compared to 73% in C1), with the disparity being most prominent in the Black/African-American demographic with a wide

[b] Supplementary information is available at: https://github.com/EpistasisLab/PSB2024_spine/

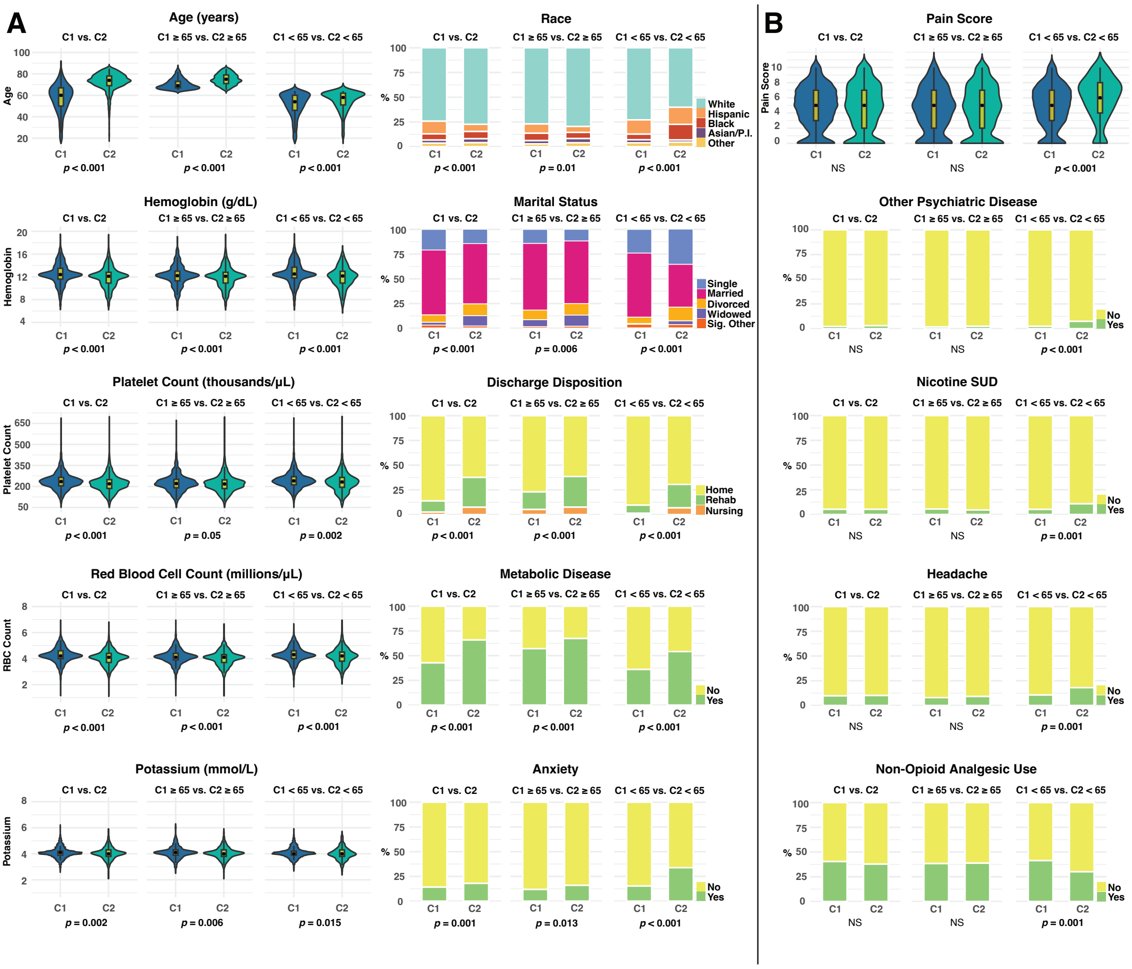
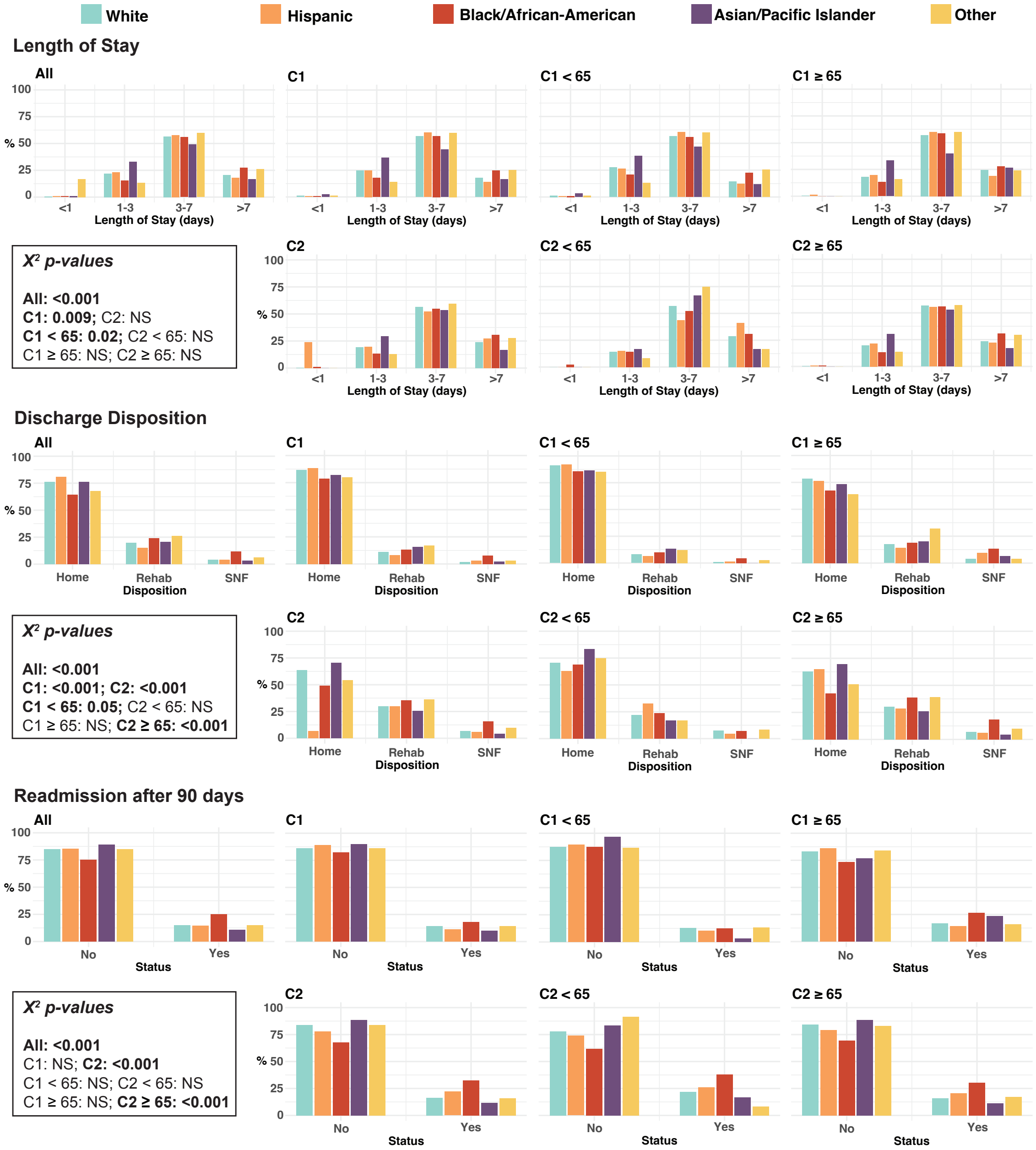


Fig. 3. Pairwise comparison results of selected features for Exp 1, 2, & 3 (C1 vs. C2; C1 ≥ 65 vs. C2 ≥ 65; C1 < 65 vs. C2 < 65) significant across all in (A), and only for Exp 3 in (B).

percentage gap of 16% vs 5.7% (Figure 3). Given a predominantly White cohort, it is important to highlight that complex ML models may inadvertently neglect pattern associations within minority classes. We recognize the importance of deeper exploration into race since our clustering model could potentially marginalize significant patterns linked to minority groups. This section further examines race-related differentiation at both cohort and cluster levels.

We observe significant differences for post-operative care outcomes (discharge disposition, length of hospital stay (LOS), and readmission rate) between race groups in multiple comparisons (Figure 4). At the entire cohort level, Blacks exhibit a higher proportion of adverse outcomes in all scenarios (see Figure 4). The "Other" group (Native American or Alaskan Native, Other, patient declined, and unknown) also demonstrates increased rates of adverse outcomes for discharge disposition and LOS. We subsequently examine the cluster and age-stratified groups to identify whether the adverse outcome over-representation in Blacks and "Other" remain independent of insurance and age. Likewise, for readmission rate and discharge disposition, the higher adverse outcome effect remains significant in C2, specifically in the 65+ subgroup. However, LOS is independent of race in C2 as adverse outcomes become



X² ***p-values***

All: <0.001
C1: 0.009; C2: NS
C1 < 65: 0.02; C2 < 65: NS
C1 ≥ 65: NS; C2 ≥ 65: NS

X² ***p-values***

All: <0.001
C1: <0.001; C2: <0.001
C1 < 65: 0.05; C2 < 65: NS
C1 ≥ 65: NS; **C2 ≥ 65: <0.001**

X² ***p-values***

All: <0.001
C1: NS; **C2: <0.001**
C1 < 65: NS; C2 < 65: NS
C1 ≥ 65: NS; **C2 ≥ 65: <0.001**

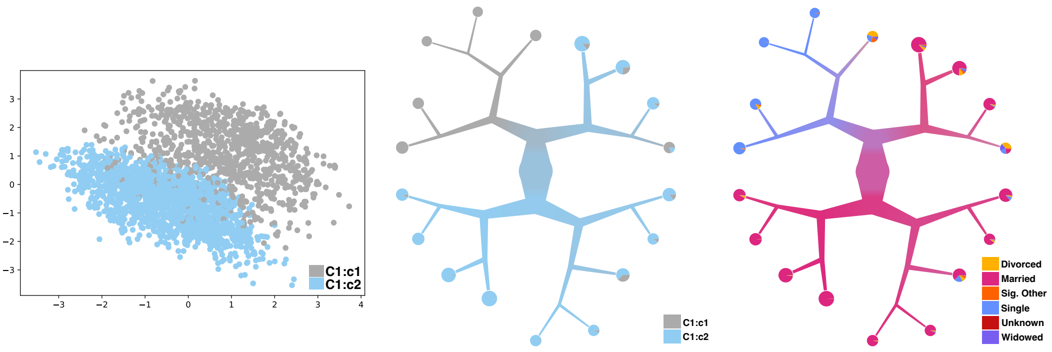
Fig. 4. Pairwise comparisons of clinical outcomes across subgroups by race.

more prominent for all groups, likely denoting a combined effect of socioeconomic disparities and advanced age. Race appears to also be an important factor in C1 with Blacks and “Other” having higher LOS (> 7 days) and discharges to other than home compared to other groups. These results, although limited due to small non-white sample sizes, indicate that race is an important discriminant of health outcomes for ESF surgery.

4.2. *Second-order clustering reveals clinical and demographic heterogeneity*

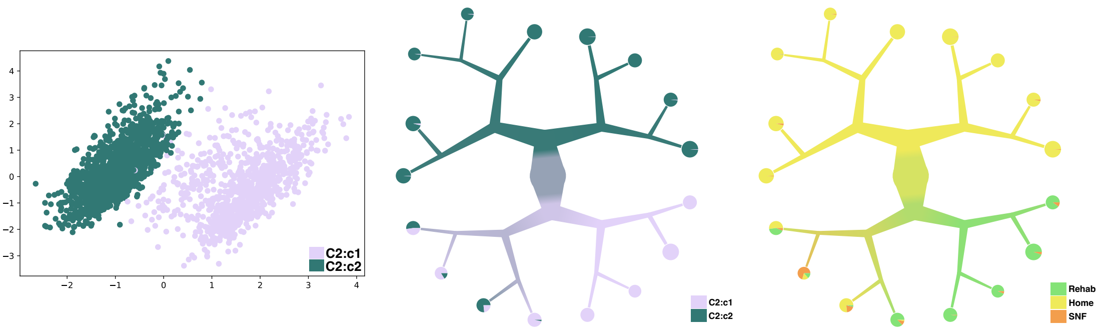
Given the overwhelmingly distinct clusters driven by socioeconomic factors, we reiterate the automated clustering on C1 and C2 separately to further examine the insurance-associated heterogeneity. This is denoted as second-order clustering (Exp 6 and 7) in Figure 2.

The top-ranking clustering results for both experiments are illustrated in Table 2B and C. We observe that in both instances, the k-means-2 result is the most optimal method. For C2, all the high-ranking algorithms unanimously identified 2-cluster solutions with minor size distribution differences. In contrast for C1, though the 2-cluster solution is the best method overall, there is more variance among the metrics. Visual inspection of ISOMAP decomposition and TMC dendrograms with cluster labels confirm that C2 clusters (C2:c1 and C2:c2) display more separation compared to C1 (C1:c1 and C1:c2) (Figures 5 and 6).



(a) ISOMAP representation (b) TMC with $k$means-2 (c) Marital status TMC

Fig. 5. Optimal clustering result on C1 subgroup:kmeans-2 optimal result.



(a) ISOMAP representation (b) TMC with $k$means-2 (c) Discharge TMC

Fig. 6. Optimal clustering result on C2 subgroup: $k$means-2 optimal result.

TPOT feature importance analysis identifies marital status as highly discriminant for C1:c1 and C1:c2 groups, and discharge disposition for C2:c1 and C2:c2, both with 100% B-Acc. For C1, Age trails with 57.4% B-Acc. For C2, LOS, hemoglobin, and readmit predicts label with 64.2%, 61.8%, and 60.9% B-Acc. respectively. This is illustrated using the TMC dendrograms overlaid with the discriminant features in Figures 5(c) and 6(c). C1:c2 consists entirely of all married patients while C1:c1 contains all others. In C2, the two clusters (C2:c1 and C2:c2) are stratified primarily by discharge disposition. C2:c1 (n = 886) consists mainly of patients discharged to rehab and skilled-nursing facilities (SNF) while C2:c2 (n = 1,476) is comprised of almost all home discharge patients (99.86%). We also observe that the second-

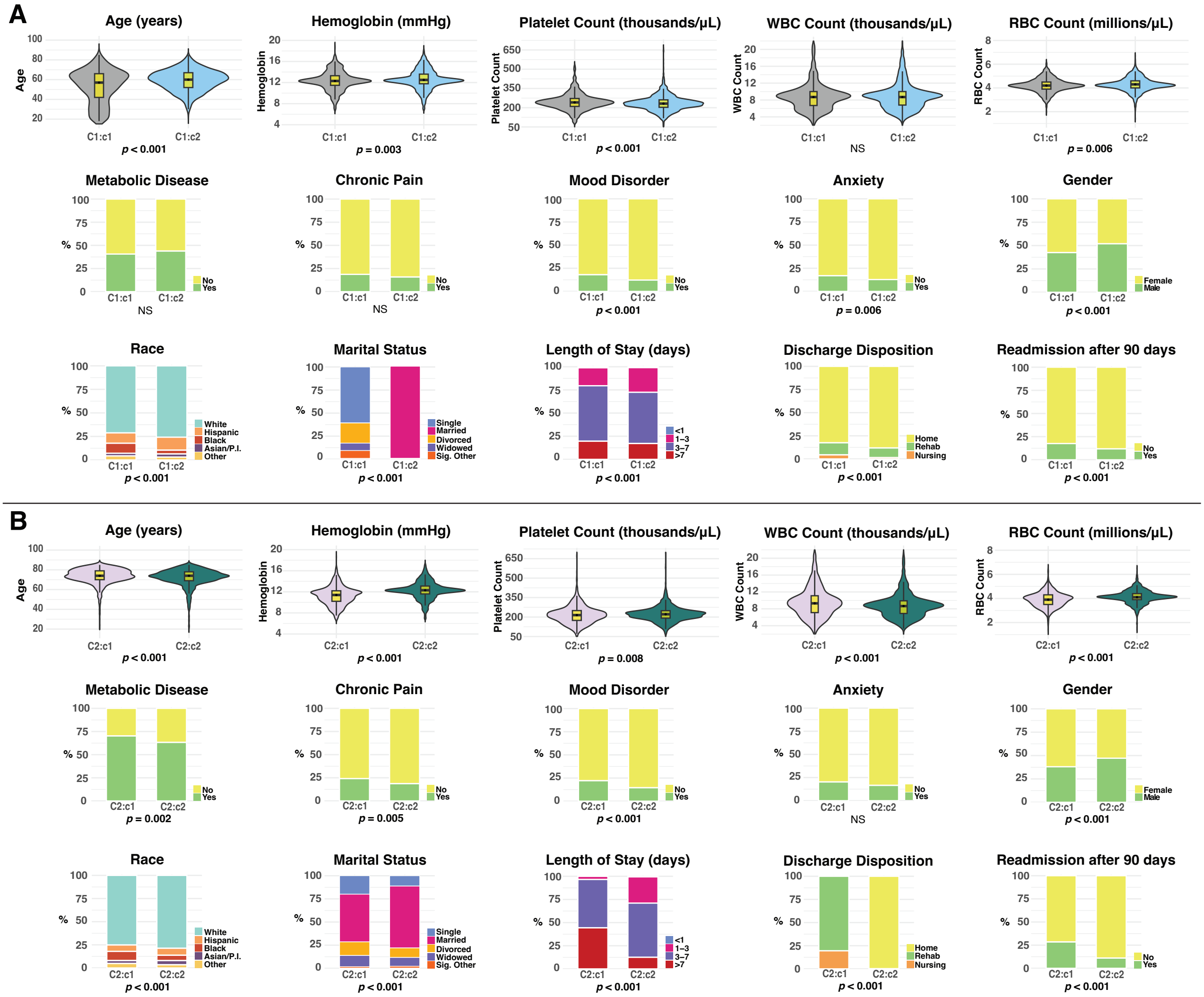


Fig. 7. Selected significant features from univariate analysis of pairwise comparisons on second-order clustering on (A) cluster C1 (Exp 6) and (B) cluster C2 (Exp 7).

order clustering yields subgroups of disproportionate sizes (large vs. small) compared to the first-order clustering.

From univariate analysis results (Figure 7), statistically significant differences are observed for both comparisons (Exp 6 and 7) for age, race, gender, discharge disposition, readmission, LOS, platelet count, RBC count, hemoglobin, BUN, creatinine, chloride, calcium, sodium, and PMH features of anxiety and mood of which selected features are illustrated in 7. Overall, we observe that C2 displays a higher level of complexity and divergence. The features that drive the C2:c1 vs. C2:c2 divergence are LOS > 7 days (44% vs 13%), readmission rate (29% vs 11%), and lower median hemoglobin values (11.4 vs. 12.3) (Figure 7B).

5. Discussion

In this study, we elaborate our commitment towards constructing equitable and unbiased ML models. Our initial intention was the development of a predictive model specific to elective spine fusion surgery, however, during the course of our investigation, we identified the necessity for deeper understanding of potential disparities present within our dataset to more accurately address clinical inquiries. The manifestation of bias within ML algorithms through data sources has been substantially highlighted in prior literature.[13,29,30] To combat this, we employ a robust automated multiple clustering approach to scrutinize our dataset for potential bias factors, prior to developing an ML model. Investigation of subpopulation structure in clinical cohorts is an important area of research and has significant implications for patient care and treatment. However, the methodologies used in most studies[15–17] are limited in that they usually implement a single clustering technique without conducting exploratory investigations of their results, potentially overlooking components driving heterogeneity. Our framework addresses these shortcomings by employing automated cluster analysis with hyperparameter tuning and a multi-metric performance score. The framework, enhanced by autoML-driven feature importance estimation along with univariate analysis, allowed us to uncover and explain drivers of population divergence. We demonstrate its capabilities in uncovering inherent patterns of heterogeneity in patients undergoing ESF, an invasive medical procedure that is associated with risks of many adverse outcomes.[11]

The cluster analysis uncovers two diverse subgroups (C1 and C2), each exhibiting unique characteristics, driven mainly by socioeconomic factors (insurance type and race). It is important to note that the entire ESF sample is almost evenly split between insurance types (54% commercial insurance). This indicates increasing equity of access as patients with medicare coverage have historically experienced limited access to certain medical procedures, including elective spinal fusion.[10] However, disheartening but not surprising, is the observed significant health disparities in the cohort driven by socioeconomic factors. Similarly, there are several recent studies[31,32] highlighting that racial minorities, and those with lower socioeconomic status, are at higher risk of adverse outcomes. The C2 subgroup contains all medicare insurance patients and is characterized by an increased proportion of minority groups compared to C1, though the overall sample is primarily White (Table 1). C2 patients have higher occurrences of non-home discharge dispositions, clinically remarkable past medical histories, especially with respect to metabolic-related diseases and anxiety, as well as clinical lab values associated with poor prognoses (Figure 3). In particular, the under 65 C2 patients (266) have significantly higher pain scores and a higher prevalence of nicotine substance abuse, headaches, other psychiatric disorders, and conditions already noted (metabolic and anxiety). These characteristics are not surprising, however, what is notable is that the socioeconomic factor of insurance overwhelms the clustering results, compelling us to adjust for it prior to characterizing the underlying heterogeneity with second-order clustering on C1 and C2 separately.

Both C1 and C2 contain one of two sub-clusters that are smaller and associated with poor health outcomes (C1:c1 and C2:c1). Interestingly, C1 is stratified by marital status with C1:c2 consisting of all married patients while C1:c1, its adverse outcome subcluster, is made up of all other marital status groups (Figure 5(c)). C2 is stratified by discharge disposition. C2:c1, its adverse outcome group, consists of almost all non-home discharged patients (99.86%) (Figure 6(c)). Despite the unique characteristics that differentiate the adverse outcome subclusters (C1:c1 and C2:c1), they share striking similarities as both are comprised of patients presenting suboptimal values of numerous labs, PMH of mood disorder, poor outcomes (LOS, discharge, and readmission), and higher proportions of minority patients (Figure 7). Though similarities exist, the proportions of patients with negative indicators of health (and their magnitudes) are greater in C2:c1 compared to C1:c1 (Figure 7). This is also true for race as C2:c1 has a higher proportion of minority patients. This aligns with the validation metrics analysis (Table 2) which indicates more separation in C2 compared to C1. In addition, C2:c1 has significantly suboptimal WBC count, PMH of metabolic and chronic pain, and use of gabapentin while C1:c1 has more prominence of PMH of anxiety, alcohol, other psychiatric disorders, nicotine use, and other SUDs (Figure 7). We acknowledge that these characteristics are probably due to a combination of social, environmental, and biological factors. However, interestingly, overall better prognoses are strongly associated with "married" status (Figure 7A).

The conspicuous racial partitioning observed at both levels of clustering highlights the importance of conducting thorough exploratory analysis and incorporation of fair algorithms in ML. The race-stratified analysis further validates findings on existing socioeconomic disparities within the ESF sample, especially for post-surgery event outcomes (Figure 4). All relatively poor outcome subgroups (C2 as a whole, under 65 age-stratified cohort in C2, C1:c1, and C2:c1) have significantly more minority patients (Figures 3,7). Interestingly, the over-representation of Blacks and "Other" are similar in both C1:c1 and C2:c1 (Blacks: ≈10% and "Other": ≈4.5% (which includes Native Americans)). This is concerning given the overall low percentage of Blacks (6.75%) and "Other" (3.72%) in the entire sample. Note that "Other" also includes self-reported race entries of "Other", "patient declined", and "Unknown", which are often associated with privacy, self-identity/profiling, and trust concerns.[33] Constructing "Other" with Native-Americans, Alaskan Natives, and individuals with no reported race is not optimal and was done due to small sample sizes. Nevertheless, identifying higher proportions of these individuals in the adverse risk clusters is likely driven by cumulative disparity factors associated with these groups. These implications are important as identifying patients with needs for specialized care could lead to substantial improvements in clinical outcomes.

Complex pattern recognition models can sometimes overlook minority groups due to imbalanced data, potentially leading to biased results and unfair outcomes.[13] Here, we showcase a framework that mitigates these issues by incorporating information about heterogeneous subgroups into the clinical risk score model. With thorough evaluation and validation, our discovery from clustering results has the potential to be actionable in clinical settings, allowing diverse groups of patients and clinicians to receive more precise estimates of treatment success and risk of developing adverse effects. This approach can be transferred to other domains that require clinical decision support. Moreover, as we observe racial and socioeconomic indicators playing key roles in explaining disproportional adverse effect distribution, it is important to continue advocating for more fair healthcare policies, especially for preventative care access. By identifying socioeconomic status and race as significant determinants of health outcomes, our two-tier approach averts a potential scenario of introducing health disparities due to algorithmic bias. We are enthusiastic about the development and deployment of our methodology in predictive modeling in clinical settings to assist surgeons and patients in real-time decision-making regarding the most efficacious ESF surgery options. These clusters could be utilized in a sampling scheme to mitigate bias in ML models aimed at predicting outcomes, by incorporating feature engineering based on the cluster labels into the model as well as exploring risk score ML models with discovered population stratification. This study presents a compelling illustration of the heterogeneity within the healthcare system and underscores the need for personalized medicine as a strategic approach to enhance healthcare and reduce health disparities. Therefore, we strongly advocate for others to employ a similar rigorous approach to data integration in order to better comprehend potential biases.

References

1. P. Rajpurkar, E. Chen, O. Banerjee and E. J. Topol, AI in health and medicine, *Nat Med* **28**, 31 (Jan 2022).
2. A. Goyal, C. Ngufor, P. Kerezoudis, B. McCutcheon, C. Storlie and M. Bydon, Can machine learning algorithms accurately predict discharge to nonhome facility and early unplanned readmissions following spinal fusion? analysis of a national surgical registry: Presented at the 2019 aans/cns section on disorders of the spine and peripheral nerves, *Journal of Neurosurgery: Spine* **31**, 568 (2019).
3. C. Krittanawong, H. U. H. Virk, S. Bangalore, Z. Wang, K. W. Johnson, R. Pinotti, H. Zhang, S. Kaplin, B. Narasimhan, T. Kitai *et al.*, Machine learning prediction in cardiovascular diseases: a meta-analysis, *Scientific reports* **10**, p. 16057 (2020).
4. U. Ahmed, G. F. Issa, M. A. Khan, S. Aftab, M. F. Khan, R. A. Said, T. M. Ghazal and M. Ahmad, Prediction of diabetes empowered with fused machine learning, *IEEE Access* **10**, 8529 (2022).
5. C. Kavitha, V. Mani, S. Srividhya, O. I. Khalaf and C. A. Tavera Romero, Early-stage alzheimer's disease prediction using machine learning models, *Frontiers in public health* **10**, p. 853294 (2022).
6. P.-Y. Tseng, Y.-T. Chen, C.-H. Wang, K.-M. Chiu, Y.-S. Peng, S.-P. Hsu, K.-L. Chen, C.-Y. Yang and O. K.-S. Lee, Prediction of the development of acute kidney injury following cardiac surgery by machine learning, *Critical care* **24**, 1 (2020).
7. B. M. Stopa, F. C. Robertson, A. V. Karhade, M. Chua, M. L. Broekman, J. H. Schwab, T. R. Smith and W. B. Gormley, Predicting nonroutine discharge after elective spine surgery: external validation of machine learning algorithms: Presented at the 2019 aans/cns joint section on disorders of the spine and peripheral nerves, *Journal of Neurosurgery: Spine* **31**, 742 (2019).
8. A. Siccoli, M. P. de Wispelaere, M. L. Schröder and V. E. Staartjes, Machine learning–based preoperative predictive analytics for lumbar spinal stenosis, *Neurosurgical Focus* **46**, p. E5 (2019).
9. S. S. Rajaee, H. W. Bae, L. E. Kanim and R. B. Delamarter, Spinal fusion in the united states: analysis of trends from 1998 to 2008, *Spine* **37**, 67 (2012).
10. D. Badin, C. Ortiz-Babilonia, F. N. Musharbash and A. Jain, Disparities in elective spine surgery for medicaid beneficiaries: a systematic review, *Global Spine Journal* **13**, 534 (2023).
11. S. J. S. Bajwa and R. Haldar, Pain management following spinal surgeries: an appraisal of the available options, *Journal of craniovertebral junction & spine* **6**, p. 105 (2015).
12. A. Finkelstein, M. Gentzkow and H. Williams, Sources of geographic variation in health care: Evidence from patient migration, *The quarterly journal of economics* **131**, 1681 (2016).
13. T. P. Pagano, R. B. Loureiro, F. V. N. Lisboa, G. O. R. Cruz, R. M. Peixoto, G. A. d. S. Guimarães, L. L. d. Santos, M. M. Araujo, M. Cruz, E. L. S. de Oliveira *et al.*, Bias and unfairness in machine learning models: a systematic literature review, *arXiv preprint arXiv:2202.08176* (2022).
14. Z. Obermeyer, B. Powers, C. Vogeli and S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, *Science* **366**, 447 (Oct 2019).

15. T. Ahmad, M. J. Pencina, P. J. Schulte, E. O'Brien, D. J. Whellan, I. L. a, D. W. Kitzman, K. L. Lee, C. M. O'Connor and G. M. Felker, Clinical implications of chronic heart failure phenotypes defined by cluster analysis, *J Am Coll Cardiol* **64**, 1765 (Oct 2014).
16. L. Marisa, A. s, A. Duval, J. Selves, M. P. Gaub, L. Vescovo, M. C. Etienne-Grimaldi, R. Schiappa, D. Guenot, M. Ayadi, S. Kirzin, M. Chazal, J. F. jou, D. Benchimol, A. Berger, A. Lagarde, E. Pencreach, F. Piard, D. Elias, Y. Parc, S. Olschwang, G. Milano, P. Laurent-Puig and V. Boige, Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value, *PLoS Med* **10**, p. e1001453 (2013).
17. K. A. e. a. Hoadley, Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer, *Cell* **173**, 291 (Apr 2018).
18. S. van Buuren and K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in R, *Journal of Statistical Software* **45**, 1 (2011).
19. K. Al-Jabery, T. Obafemi-Ajayi, G. Olbricht and D. Wunsch, *"Computational Learning Approaches to Data Analytics in Biomedical Applications"* (Academic Press, 2019).
20. T. Ronan, S. Anastasio, Z. Qi, R. Sloutsky, K. M. Naegle and P. H. S. V. Tavares, Openensembles: a python resource for ensemble clustering, *The Journal of Machine Learning Research* **19**, 956 (2018).
21. D. Yeboah, L. Steinmeister, D. B. Hier, B. Hadi, D. C. Wunsch, G. R. Olbricht and T. Obafemi-Ajayi, An explainable and statistically validated ensemble clustering model applied to the identification of traumatic brain injury subgroups, *IEEE Access* **8**, 180690 (2020).
22. G. W. Schwartz, Y. Zhou, J. Petrovic, M. Fasolino, L. Xu, S. M. Shaffer, W. S. Pear, G. Vahedi and R. B. Faryabi, TooManyCells identifies and visualizes relationships of single-cell clades, *Nat Methods* **17**, 405 (Apr 2020).
23. T. en, K. Nowell, K. E. Bodner and T. Obafemi-Ajayi, Ensemble validation paradigm for intelligent data analysis in autism spectrum disorders, in *2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2018.
24. Y. Liu, Z. Li, H. Xiong, X. Gao and J. Wu, Understanding of internal clustering validation measures, in *2010 IEEE international conference on data mining*, 2010.
25. R. S. Olson, R. J. Urbanowicz, P. C. Andrews, N. A. Lavender, L. C. Kidd and J. H. Moore, Automating biomedical data science through tree-based pipeline optimization, in *Applications of Evolutionary Computation*, eds. G. Squillero and P. Burelli (Springer International Publishing, Cham, 2016).
26. S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, From local explanations to global understanding with explainable ai for trees, *Nature Machine Intelligence* **2**, 2522 (2020).
27. L. Breiman, Random forests, *Machine learning* **45**, 5 (2001).
28. J. B. Tenenbaum, V. de Silva and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* **290**, 2319 (Dec 2000).
29. W. Sun, O. Nasraoui and P. Shafto, Evolution and impact of bias in human and machine learning algorithm interaction, *Plos one* **15**, p. e0235502 (2020).
30. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, A survey on bias and fairness in machine learning, *ACM computing surveys (CSUR)* **54**, 1 (2021).
31. L. Wang, N. A. Berger, D. C. Kaelber, P. B. Davis, N. D. Volkow and R. Xu, Covid infection rates, clinical outcomes, and racial/ethnic and gender disparities before and after omicron emerged in the us, *medRxiv* (2022).
32. D. Quan, L. Luna Wong, A. Shallal, R. Madan, A. Hamdan, H. Ahdi, A. Daneshvar, M. Mahajan, M. Nasereldin, M. Van Harn *et al.*, Impact of race and socioeconomic status on outcomes in patients hospitalized with covid-19, *Journal of general internal medicine* **36**, 1302 (2021).

33. S. J. Hong, B. Drake, M. Goodman and K. A. Kaphingst, Race, trust in doctors, privacy concerns, and consent preferences for biobanks, *Health communication* **35**, 1219 (2020).