Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Operationalizing Trustworthy AI: A Scalable Framework for Granular Error Analysis and Bias Mitigation in Healthcare Models

Andrea Corvaglia
Vita-Salute San Raffaele University
Marco Montagna
Vita-Salute San Raffaele University
Davide Vignale
Vita-Salute San Raffaele University
Anna Palmisano
Vita-Salute San Raffaele University
Francesco Pisu
Vita-Salute San Raffaele University
Alberto Colombo
Vita-Salute San Raffaele University
Marta Liberotti
Vita-Salute San Raffaele University
Clara Soggetti
Vita-Salute San Raffaele University
Marco Denti
Vita-Salute San Raffaele University
Patrizia Rovere Querini
Vita-Salute San Raffaele University
Carlo Tacchetti
Vita-Salute San Raffaele University
Antonio Esposito
Vita-Salute San Raffaele University
Alberto Traverso
traverso.alberto@hsr.it

Vita-Salute San Raffaele University

Article

Keywords:

Posted Date: July 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9926339/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Healthcare Artificial Intelligence models often fail to generalize safely because standard aggregate performance metrics act as "statistical masks," obscuring severe algorithmic failures within specific patient subgroups. To address this "translation gap," we present a scalable framework within the San Raffaele Artificial Intelligence Centre platform that utilizes an Error Analysis Decision Tree to automatically discover cohorts with concentrated error rates. We introduce two methodological innovations: a probability-preserving wrapper that extends Responsible Artificial Intelligence diagnostics to complex survival models, and a feature space expansion technique enabling fairness auditing with demographic variables intentionally excluded from model inference. Validated on two clinical use cases —predicting short-term mortality in multimorbid elderly patients and one-year mortality post-Transcatheter Aortic Valve Implantation—the framework successfully unmasked divergent, hidden failure modes. By operationalizing active Algorithmic Vigilance, this approach translates abstract fairness principles into verifiable engineering objectives, directly addressing stringent compliance mandates like the European Union Artificial Intelligence Act.

INTRODUCTION

Artificial Intelligence (AI) and its branch, Machine Learning (ML), are poised to shift clinical practice from population-based guidelines to truly personalized medicine (Rajkomar et al., 2019). However, as Clinical Decision Support Systems (CDSSs) are deployed, a critical "translation gap" has emerged between *in silico* performance and *in vivo* reliability. AI models are overwhelmingly trained on retrospective Real-World Data (RWD), which are defined by FDA as “data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources” and not by clinical trials (Khaowroongrueng et al., 2025). As such, rather than neutral reflections of biology, RWD often act as a "socio-technical artefact" laden with selection biases and historical inequities. Without sufficient scrutiny, models risk failing to generalize and actively automating the disparities they were intended to resolve (Rajkomar et al., 2018).

This translational failure frequently stems from validation methodologies that rely on aggregate metrics (e.g., Accuracy, Area Under the Curve-AUC). These metrics act as "statistical masks," obscuring algorithmic failures within specific patient subgroups (Liu et al., 2019). Landmark studies vividly illustrate this: Obermeyer et al. demonstrated severe racial bias where Black patients were considerably sicker than White patients at identical risk scores (Obermeyer et al., 2019); Seyyed-Kalantari et al. revealed systemic socioeconomic under-diagnosis (Seyyed-Kalantari et al., 2021); and Oakden-Rayner et al. identified "hidden stratification" where models fail on unlabeled, clinically meaningful subsets (Oakden-Rayner et al., 2020).

Crucially, as Otokiti et al reported, model failures are often intersectional (e.g., error rates spiking specifically for darker-skinned females) (Otokiti et al., 2025). Because patients exist at the intersection of

hundreds of clinical and demographic variables, manual validation is mathematically and computationally infeasible (Nushi et al., 2018).

Even advanced metrics like Decision Curve Analysis (DCA) are insufficient (Vickers and Elkin, 2006). DCA assumes population-wide calibration and calculates an averaged Net Benefit, which can mathematically mask scenarios where a model generates utility for the majority but inflicts harm (negative Net Benefit) on a minority subgroup (Paulus and Kent, 2020) (On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative et al., 2019). Furthermore, manually slicing data to audit False Positive and False Negative rates is constrained by human cognitive limits. Consequently, the search for error patterns must be algorithmic rather than manual.

Granular auditing is no longer just an ethical imperative; it is a stringent legal requirement. The EU Artificial Intelligence Act (AI Act) and the Medical Device Regulation (MDR 2017/745) (European Parliament. Directorate General for Parliamentary Research Services., 2022) mandate that "High-Risk" AI systems demonstrate consistent performance across different user groups and actively mitigate bias. Failing to detect "hidden stratification" constitutes a direct regulatory non-compliance risk. However, a significant "compliance gap" remains. Regulatory frameworks call for fairness and robustness, but do not specify the algorithmic methods needed to detect non-linear failure modes in high-dimensional feature spaces.

Automated, data-driven error analysis is essential to bridge the "implementation gap" of Responsible AI (*Ethics and Governance of Artificial Intelligence for Health*, 2021), translating normative values into verifiable technical metrics. Aligning with the "Identify-Diagnose-Mitigate" lifecycle, Algorithmic Vigilance shifts validation from passive "black box" monitoring to active auditing (Van Genderen et al., 2025).

In this manuscript, we present a framework for data-driven model failure discovery, utilizing the Error Analysis Decision Tree to automatically partition the input space and identify cohorts that maximize error rates. Beyond describing the integration of the Microsoft Responsible AI toolbox within the S-RACE (San Raffaele AI centre) platform (Traverso et al., 2026), we provide empirical evidence from concrete clinical use cases. We detail methodological solutions developed to overcome key real-world data barriers, including auditing fairness when demographic attributes are excluded from model inference and extending Responsible AI diagnostics to survival models, thereby translating abstract principles of fairness into measurable and auditable engineering objectives.

RESULTS

We applied the Responsible AI error analysis workflow to two real clinical use cases integrated in our S-RACE platform, using held-out evaluation cohorts to evaluate aggregated performance metrics. Across both clinical examples, the analysis was explicitly positioned as a debugging and governance step in the model development sequence: (i) discover concentrated failure modes via error-focused cohort splitting, (ii) characterize error type and score behavior within those cohorts, and (iii) translate findings into concrete hypotheses for further data/model refinement.

Practical Example I: in-hospital or post-discharge mortality in complex multimorbid elderly patients

Ageing of populations and the rising prevalence of multimorbidity complicate acute decision-making in internal medicine (Tinetti et al., 2004). Unlike oncology, this field lacks robust, generalizable prognostic tools to guide management choices between an aggressive approach and a more conservative or palliative one. These decisions frequently fall within a clinical “grey zone” where the balance between potential benefit and harm is uncertain (Walling et al., 2010). Many elderly patients with advanced, non-malignant chronic diseases follow fluctuating illness trajectories characterized by periods of functional decline interspersed with acute life-threatening exacerbations. These patterns make it difficult to anticipate whether a given episode represents a reversible event or part of a progressive end-of-life trajectory, yet such anticipation is crucial for aligning interventions with patient goals and avoiding treatments unlikely to provide meaningful benefit. Existing prognostic scores typically estimate long-term (e.g., one-year) mortality or rely on administrative datasets that poorly represent the frail, complex multimorbid population commonly admitted to internal medicine wards (Tinetti et al., 2004) (Tinetti and Fried, 2004). Therefore, moving beyond a disease-centric view toward a multidimensional, real-world data–driven approach is essential to capture vulnerability, functional status, organ reserve, and evolving care preferences. Such an approach may ultimately support more nuanced, timely, and person-centered decisions in these clinically challenging scenarios. To address this gap, the prospective MED-Cli trial (NCT05780099, registration date 24-02-2026) at IRCCS San Raffaele Hospital systematically collects sociodemographic, clinical, functional, laboratory, and therapeutic longitudinal real-world inpatient data. These data are ingested into the S-RACE platform to develop ML instruments predicting short-term disease outcomes. We developed a Cox proportional hazards model to predict in-hospital or 90-day post-discharge mortality, using a parsimonious set of multidimensional clinical features.

- Cohorts: 325 admission episodes (training/validation) and 157 episodes (temporal testing).
- Predictive Features (4): age, sex, Braden pressure ulcer risk score (fitness proxy), and dysphagia presence/absence.
- Results: The model achieved a mean Concordance Index (C-index) of 0.72 (95% CI: 0.57–0.87) across 10-fold nested cross-validation, and 0.66 in the independent temporal validation cohort. Applying the model effectively stratified the temporal testing cohort into low-risk (5.3%), intermediate-risk (26.8%), and high-risk (58.3%) mortality groups.

Med-CLI: survival-aware auditing at a 90-day horizon reveals two distinct high-error phenotypes

To demonstrate survival-compatible auditing within a clinical risk workflow, we analyzed the Med-CLI dataset using a time-to-event composite endpoint at a 90-day horizon on the evaluable cohort (n = 118).

The resulting error tree identified two primary high-error cohorts driven by patient performance and age (Fig. 1A):

- Cohort 1 (Low fitness): Braden Score ≤ 14.50 (15 misclassifications out of 29).
- Cohort 2 (Higher fitness but Older-old): Braden Score > 14.50 and Age > 84.50 (9 misclassifications out of 20).

Conversely, the adjacent subgroup within the higher-fitness branch (Braden Score > 14.50 and Age ≤ 84.50) exhibited a substantially lower error rate (7 misclassifications out of 69). A clinically interpretable split can be identified, whereby the model is particularly bad at predicting outcomes for the "older-old" patients within an otherwise similar performance stratum, representing a possible lack of fairness for this age group.

While the aggregate confusion matrix summarizes overall performance at the selected operating point, isolating the two high-error cohorts reveals opposing error profiles that are otherwise masked.

Cohort-specific confusion matrices demonstrate that the older-old subgroup (Age > 84.50) and the low-fitness subgroup (Braden ≤ 14.50) concentrate entirely distinct types of misclassifications. Comparing error decomposition metrics, specifically the false positive rate, false negative rate, and selection rate, across the full evaluation set and each subgroup confirms this divergence: the former subgroup (cohort 1) has a markedly high false negative rate, while the latter subgroup (cohort 2) has comparable false positive and false negative rates (Fig. 1B-C). This highlights that although both cohorts are classified as "high error," their fundamental modes of algorithmic failure are different.

These divergent error modes are further reflected in the model's score behavior. The probability-like risk score distributions shift in cohort-specific ways relative to the overall evaluation set, suggesting that miscalibration and separability issues are not uniform across patient strata.

To translate these descriptive findings into actionable debugging hypotheses, we examined cohort-specific feature contributions and marginal effects:

- Feature Importance: The relative contributions of age and Braden score differ significantly between the two failure cohorts and the overall dataset.
- Marginal Effects: The distribution of per-feature effects exhibits cohort-dependent shifts, pointing to distinct underlying mechanisms of failure.
- Partial Dependence Plots: contextualize how risk estimates vary with age and Braden scores when averaging over other covariates.

Collectively, this evidence—spanning error types, score distributions, and feature effects—supports a concrete refinement hypothesis. The model likely requires additional clinical descriptors to effectively discriminate events in (i) very old patients with higher Braden scores, and (ii) patients with low Braden scores. This insight directly motivates targeted feature enrichment and subsequent external validation.

Practical example II: Predicting 1-Year Mortality After Transcatheter Aortic Valve Implantation (TAVI)

Transcatheter Aortic Valve Implantation (TAVI) is the treatment of choice for severe symptomatic aortic stenosis in elderly or high surgical risk patients (Praz et al., 2025). However, 1 year mortality after TAVI can exceed 10–20% in real-word cohorts (Tamburino et al., 2015) (Ferry et al., 2025), highlighting a substantial proportion of possible futile procedures. Standard surgical risk estimators, such as the STS score and EuroSCORE II, are not tailored for percutaneous procedures and tend to overestimate risk (Webb et al., 2022). Consequently, accurate prognostic assessment for these complex, multimorbid patients remains a major clinical challenge.

To address this unmet need, a retrospective, single-center observational study was conducted on consecutive participants undergoing TAVI. The objective was to predict 1-year all-cause mortality by leveraging pre-procedural cardiac CT data—a rich but currently underutilized resource in prognostic prediction.

A reduced-CT machine learning model based on Extremely Randomized Trees was derived from a pre-procedural cardiac CT feature set using Boruta feature selection within a nested validation framework, followed by permutation-feature-importance-driven reduction to balance predictive performance and parsimony.

Cohorts: 412 participants in the derivation cohort and 135 participants in the internal temporal validation cohort.

Predictive features (N = 8): coronary artery disease/calcification category, LVOT (left ventricular outflow tract) mean diameter, right atrial end-diastolic volume, indexed aortic valve calcium score, annulus mean diameter, extracellular volume fraction, heart rate during cardiac CT, and ascending aorta mean diameter.

Results: The reduced-CT model achieved an AUC of 0.72 (95% CI: 0.65–0.79) in the internal test set and maintained an AUC of 0.72 (95% CI: 0.58–0.86) in the internal temporal validation cohort. When predefined tertile cut-offs were applied to the validation cohort, the model stratified participants into low-, intermediate-, and high-risk groups with 1-year mortality rates of 3%, 10%, and 26%, respectively.

Practical Example II provides a complementary demonstration in a standard binary classification setting, using the eight-feature reduced-CT model for 1-year mortality after TAVI. The predictive model was based exclusively on CT-derived variables; no demographic or clinical variables were used for inference. Sex was therefore included only as an audit-only example variable to validate the feature-space expansion strategy, demonstrating that clinically relevant metadata excluded from model training can still be used for post hoc subgroup auditing. In the held-out internal temporal validation cohort (n = 135), aggregate performance at the selected operating point was modest (precision 0.192, recall 0.385, false positive rate 0.172, false negative rate 0.615), with 29/135 misclassifications. This global summary masked substantial local heterogeneity.

As shown in Fig. 2A, error cohort splitting identified two CT-defined regions with markedly different error burdens. A low-error subgroup was defined by *LVOT_mean_diameter* ≤ 24.35 and *Annulus_mean_diameter* ≤ 24.45 (n = 59), with 2/59 misclassifications (3.39%) and 6.90% of the total error burden. In contrast, a high-error subgroup was defined by *LVOT_mean_diameter* > 24.35 and CACS > 2 (tree node threshold: 2.50) (n = 33), with 20/33 misclassifications (60.61%) and 68.97% of all errors in the validation cohort. Thus, model failure was concentrated in a compact region of the CT feature space rather than being uniformly distributed across the population.

Cohort-level metrics further characterized these local failure modes (Fig. 2C). The low-error cohort generated no positive predictions (precision not applicable, recall 0, FPR 0, FNR 1), whereas the high-error cohort showed poor precision (0.105), elevated FPR (0.607), and residual missed events (recall 0.400, FNR 0.600). Predicted probability distributions were consistent with this pattern, showing a rightward shift of scores in the high-error cohort relative to the overall validation cohort and the low-error cohort (Fig. 2B).

Sex was then evaluated as an audit-only attribute (n = 68 and n = 67 across the two encoded strata). At the selected operating point, the two strata differed in accuracy (0.926 vs 0.642), FPR (0.046 vs 0.316), and selection rate (0.059 vs 0.328), while FNR was more similar (0.667 vs 0.600). These differences should not be interpreted as evidence of clinical unfairness, but as findings that warrant further subgroup-specific investigation. They nevertheless demonstrate that potential demographic performance differences can be surfaced even when sex is not used by the predictive model.

Figure 2: Error analysis and fairness assessment of the TAVI mortality prediction model.

(A) Error analysis trees for the internal temporal validation cohort (n = 135), generated for the eight-feature reduced-CT model. The root node summarizes 29/135 misclassifications. A low-error region defined by LVOT_mean_diameter ≤ 24.35 and Annulus_mean_diameter ≤ 24.45 contains 2 errors among 59 cases (error rate, 3.39%; error coverage, 6.90%). A high-error region defined by LVOT_mean_diameter > 24.35 and CACS > 2.50 contains 20 errors among 33 cases (error rate, 60.61%; error coverage, 68.97%).(B) Distribution of predicted probabilities across the overall validation cohort and the error-defined subgroups, showing a rightward shift in the high-error cohort.(C) Cohort-level classification metrics at the selected operating point, reporting sample size, precision, recall, false-positive rate, and false-negative rate for the overall cohort and the two error-defined subgroups. (D) Audit-only sex assessment reporting group-wise performance and derived fairness metrics. Sex was not used by the predictive model and was included only as an example variable for post hoc subgroup auditing. Observed differences in accuracy, false-positive rate, and selection rate do not establish unfairness, but identify subgroup-specific patterns that warrant further investigation.

DISCUSSION

The deployment of Artificial Intelligence (AI) in healthcare holds the transformative potential to shift clinical practice from population-based guidelines to truly personalized medicine. However, a critical

"translation gap" persists between theoretical *in silico* performance and *in vivo* clinical reliability. Because models are predominantly trained on retrospective Real-World Data (RWD), which acts as a socio-technical artefact laden with historical inequities rather than a neutral reflection of biology, there is a profound risk of automating and scaling the very health disparities they are intended to resolve. In this study, we demonstrate a scalable, automated framework for granular error analysis, operationalized within the S-RACE platform, which transforms the abstract ethical principles of Responsible AI (RAI) into verifiable, regulatory-compliant engineering objectives.

A central theme in contemporary health equity literature is the inadequacy of global performance metrics (e.g., accuracy, AUC, or standard decision curve analysis), which act as "statistical masks" that obscure severe algorithmic failures within minority or vulnerable patient subgroups. While foundational studies by Obermeyer et al. and Seyyed-Kalantari et al. have successfully exposed racial and socioeconomic biases, the prevailing methodological approach in the literature relies heavily on manual, *a priori* subgroup specification (e.g., manually slicing data by race or gender).

Because patients exist at the intersection of hundreds of clinical and demographic variables, this manual validation is mathematically and computationally infeasible, leaving models highly vulnerable to "hidden stratification". Furthermore, manually slicing data is highly susceptible to Simpson's Paradox (a trend appearing in different groups of data reverses or disappears when the groups are combined). Our framework diverges from these traditional methods by utilizing an unsupervised, data-driven surrogate model methodology. By fitting a LightGBM-based Error Analysis Decision Tree, the system automatically partitions the input space to identify cohorts that maximize error rates without requiring human evaluators to manually hypothesize every blind spot.

This automated discovery was validated in our Med-CLI survival model, where the tree successfully identified two distinct, high-error phenotypes: low-fitness patients (Braden Score ≤ 14.50) versus "older-old" patients with higher fitness scores (Braden Score > 14.50 and Age > 84.50). While aggregate metrics summarised overall performance at a selected operating point, isolating these cohorts revealed opposing error profiles. Cohort-specific confusion matrices and error decomposition metrics (false positive rate, false negative rate, and selection rate) confirmed that these subgroups concentrate entirely distinct types of misclassifications. This indicates divergent underlying failure mechanisms that would be completely invisible in a population-averaged evaluation or a basic demographic data slice, directly unmasking intersectional bias.

Another significant contribution of this work to the field of equitable AI is the methodological resolution of the tension between clinical parsimony and fairness auditing. In clinical machine learning, developers frequently exclude ethically and demographically sensitive attributes (e.g., sex, socioeconomic status) from the predictive model to prevent direct bias—a practice often termed "fairness through unawareness". However, standard RAI tools restrict subgroup analysis strictly to the features used by the predictive model, rendering these unmodelled demographic variables invisible during error audits and limiting the assessment of potential unfairness.

Our feature space expansion methodology overcomes this barrier. By wrapping the trained model in a pipeline with a tailored “ColumnTransformer”, we drop all non-model variables immediately prior to inference, leaving the predictive function strictly unchanged. Simultaneously, we propagate an extended set of demographic metadata directly into the RAI framework. This decoupling of inference from analysis enables the framework to leverage unmodelled variables to stratify errors without risking information leakage.

We demonstrated this critical capability in the TAVI clinical use case, where the predictive model relied exclusively on eight pre-procedural CT-derived variables and therefore remained naturally blind to demographic attributes. By evaluating sex as an audit-only attribute, the framework showed how demographic variables excluded from model inference can still be used during validation to assess subgroup-specific performance. In this internal validation setting, differences emerged in accuracy, false-positive rate, and selection rate across sex strata, while false-negative rates were comparatively similar. These findings do not establish clinical unfairness, but may guide subsequent model refinement, threshold adjustment, subgroup review, or other fairness-oriented mitigation strategies.

While most published fairness frameworks focus strictly on standard classification, time-to-event endpoints are clinically vital but fall outside the native scope of standard classification-based RAI dashboards. Our methodology bridges this gap by operationalizing survival prediction into horizon-specific binary endpoints (e.g., a 90-day horizon). By employing a probability-preserving wrapper that delegates inference to the underlying survival model and applies a strictly monotonic transformation, we standardize outputs to the [0, 1] interval. This ensures that the RAI error tree logic directly mirrors clinical risk stratification, extending state-of-the-art fairness auditing tools to complex survival models.

The imperative for granular auditing is no longer just an ethical "nice-to-have"; it is a stringent legal requirement mandated by the AI Act and the MDR 2017/745, which require active bias mitigation and consistent performance across user groups. Failing to detect hidden stratification constitutes a direct regulatory non-compliance risk. Our approach addresses this compliance gap by shifting validation from passive "black box" monitoring to active Algorithmic Vigilance.

While our framework successfully translates descriptive findings into actionable debugging hypotheses, such as identifying the need for additional clinical descriptors in the Med-CLI model, this study is currently based on retrospective validation within internal cohorts. Future research must evaluate how these automated error trees perform across multi-institutional, federated datasets to assess geographic and demographic shifts in error cohorts. Ultimately, embedding data-driven failure discovery into routine clinical workflows is an essential step towards ensuring that clinical AI actively points out, rather than scales, systemic health inequities.

Within the model development workflow, this type of analysis can be used during evaluation to examine where model errors concentrate beyond aggregate performance metrics. In internal development settings, as illustrated by the Med-CLI example, such diagnostics can help determine whether further model refinement is warranted. In temporal or external validation, as in the TAVI case, the same analysis

helps contextualize generalization by identifying cohorts where model behavior diverges from the overall summary metrics. Two practical adaptations support its use in common clinical settings: horizon-based auditing for time-to-event models and error stratification using audit-only variables not included in the predictive model.

Despite these methodological advancements, our work has several limitations that must be acknowledged. First, this study is currently based on retrospective validation within internal cohorts. Future research must evaluate how these automated error trees perform across multi-institutional, federated datasets to assess geographic and demographic shifts in error cohorts. Second, the fairness assessments and error metrics generated by the system should be interpreted as an auditing readout tied to the selected operating point and cohort size, rather than as a definitive statement of clinical unfairness. Instead, they provide a concrete mechanism to surface and prioritize follow-up analyses, such as threshold sensitivity checks, prevalence differences, or clinically meaningful subgroup reviews. Third, the underlying Error Analysis Decision Tree is not intended to approximate the original predictive model or its decision boundary. Instead, it functions as a dedicated cohort discovery mechanism optimized to surface human-interpretable data partitions that stratify error prevalence. Finally, when computing specific diagnostic metrics, such as global feature-error association scores on large datasets, the computation may need to be performed on a capped random subsample to maintain computational tractability.

In conclusion, the transition from population-based medicine to AI-driven personalization requires more than high aggregate accuracy; it demands a granular understanding of where and why models fail. As demonstrated through the Med-CLI and TAVI use cases, global metrics often mask significant "hidden stratification" that disproportionately affects vulnerable subgroups, such as the "older-old" or those with specific frailty phenotypes. Our framework, integrated within the S-RACE platform, bridges the gap between ethical theory and clinical engineering. By automating the discovery of error-prone cohorts and enabling fairness auditing on "audit-only" variables, we provide a pathway for clinicians and developers to move beyond "fairness through unawareness." This shift toward active Algorithmic Vigilance is not merely a technical optimization but a regulatory necessity under the emerging landscape of the EU AI Act and MDR 2017/745. Ultimately, the systematic identification of these "blind spots" is a prerequisite for building clinical AI that promotes health equity rather than automating historical disparities.

METHODS

The S-RACE platform

To operationalize these advanced Responsible AI (RAI) capabilities within routine clinical workflows, we developed the S-RACE (San Raffaele AI Centre) platform [Traverso et al., DOI: 10.1038/s41746-025-02132-w].

This end-to-end ecosystem bridges the gap between raw clinical data and trustworthy AI through three core pillars:

- Universal Data Platform: Ensures regulatory-grade data acquisition and curation.
- Clinician AI Hub: Facilitates collaborative cohort design.
- Data Science Lab: Embeds the Error Analysis Decision Tree directly into the model development lifecycle.

The Error Analysis Framework and Target Construction

A cornerstone of this framework is the post-hoc Error Analysis module within the Azure ML RAI Dashboard. It employs a data-driven surrogate modelling approach to identify and characterize structured regions of the input space where a trained model underperforms.

For binary classification, the module converts standard model performance into an explicit supervision signal by defining a per-instance binary error indicator, E. This recasts model debugging as a secondary supervised learning problem. E is computed deterministically: if the predicted class label $M(X)$ disagrees with the ground-truth label Y ($M(X) \neq Y$), then $E = 1$ (incorrect). If $M(X) = Y$, then $E = 0$ (correct).

Cohort Discovery via Surrogate Decision Trees

To identify subgroups with concentrated algorithmic failure, the module fits an interpretable surrogate decision tree to model the conditional error surface $P(error \mid X)$, using original input features as candidate split variables.

This implementation relies on LightGBM constrained to a single tree estimator, ensuring discovered cohorts correspond to simple feature-level conjunctions (e.g., Age > 75 AND Creatinine > 1.2). Categorical features are handled natively, ensuring splits respect semantic categories rather than numeric encodings. Resolution and stability are governed by three hyperparameters (*max_depth, num_leaves, and min_child_samples*), which prevent combinatorial growth and the generation of statistically noisy, undersized cohorts.

The dashboard visualizes the terminal leaves—representing distinct patient cohorts—using two critical metrics:

- Error Rate: Represented by node color intensity, indicating the proportion of instances within that specific cohort where the model failed.
- Error Coverage: Represented by node fill rate, indicating the proportion of *all* dataset errors contained within this cohort, helping distinguish rare edge cases from systemic population-level failures.

Exploratory Error Slicing and Global Diagnostics

Complementing the tree, a matrix-based heatmap summarizes error prevalence across one or two user-selected features (binned continuously or aggregated categorically). This enables hypothesis-driven exploration of interaction effects without strictly hierarchical constraints. Additionally, the module computes global feature-error association scores (estimating mutual information between each feature and *E)* to contextualize which variables globally drive error occurrence.

Methodological Extensions for Clinical Reality

Standard RAI tools are designed primarily for generic classification tasks and assume that the variables available for auditing coincide with the variables used by the predictive model. In clinical machine learning this assumption often fails. We therefore developed two methodological extensions that adapt the framework to common constraints encountered in real-world clinical pipelines.

Extension 1: Expanding the Feature Space Beyond Model Inputs

Standard modules restrict subgroup analysis to features explicitly used by the predictive model. However, clinical models often rely on aggressive feature selection to control overfitting, maintain model parsimony, and preserve interpretability and usability in clinical settings. As a result, ethically or clinically relevant attributes such as demographic variables or comorbidities may be excluded from the predictive model and therefore remain invisible to standard RAI audits.

To resolve this, we introduced a feature space expansion step. The trained model is wrapped in a *scikit-learn* Pipeline containing a *ColumnTransformer* configured via the *feature_names_in_* attribute. This transformer drops all non-model variables immediately prior to inference, leaving the predictive function strictly unchanged. Simultaneously, the pipeline accepts an extended set of metadata, passing it directly to the RAI framework. By decoupling inference from analysis, the module can leverage unmodelled variables to stratify errors without risking information leakage.

Extension 2: Adapting the Framework to Survival Models

Time-to-event endpoints are clinically vital but fall outside the native scope of standard classification-based RAI dashboards. Survival models are widely used in medicine because they explicitly handle censored observations and varying follow-up times, while modelling risks that evolve over time and may be clinically relevant at different horizons. We operationalized survival prediction by assessing the risk of an event by a clinically meaningful time horizon *T* (e.g., 6, 12, or 18 months).

- Horizon-Specific Binary Endpoint: Each patient's time-to-event t_i and event indicator Δ_i are transformed into a binary label for horizon T. To prevent outcome ambiguity, only patients with an observable status at T are retained, namely those who experienced the event $(\Delta_i = 1)$ or those known to remain event-free beyond $T\,(t_i > T)$.

- Probability-Preserving Wrapper: The survival model is wrapped in a scikit-learn-compatible classifier. The wrapper delegates inference to the survival model to obtain a continuous risk score r_i, which is mapped to a bounded probability p_i via a strictly monotonic transformation: $p_i = \frac{r_i}{r_i + \tau}$
- Threshold Semantics: Here, τ is a positive scaling parameter. This standardises outputs to the [0, 1] interval while preserving relative risk ordering. A probability cutoff of 0.5 means the predicted risk equals the scaling parameter ($r_i = \tau$). Users can adjust decision thresholds to reflect their intended operating point (e.g., favouring high positive or negative predictive values), ensuring RAI error tree logic directly mirrors clinical risk stratification at horizon *T*.

Although operationalized here through the Azure ML Responsible AI Dashboard, the workflow builds on Microsoft's open-source Responsible AI ecosystem. The underlying error-analysis logic—error-indicator construction, surrogate decision-tree cohort discovery, and cohort-level metric computation—can therefore be reproduced outside Azure in standalone Python workflows, while the integrated dashboard interface remains specific to the Azure/S-RACE implementation.

Declarations

DATA AVAILABILITY

Data can be made available under request. We can grant access to the S-RACE platform to the to access the data within the secure environment of our Cloud, but without the possibility of downloading the data locally.

CODE AVAILABILITY

The custom code developed for this study is publicly available in the GitHub repository AI-UniSR/responsible-ai-dashboard-template and archived on Zenodo as version v1.0 with DOI 10.5281/zenodo.19889328. The repository includes the feature-space expansion utilities, survival-to-classification wrapper, model-preparation components, and Azure ML pipeline templates used to generate the Responsible AI analyses. The code is released under the MIT License.

AUTHOR CONTRIBUTIONS

AC, MM: Design, Execution, analysis, writing

DV, AP, FP, AC, ML: Design, analysis, reviewing,

CT, AE, AT, PRQ, MD: reviewing, supervision.

All the authors have read and approved the manuscript.

ACKNOWLEDGEMENTS

This work was supported by the Italian Ministry of Research, Complementary Actions to the NRRP “D34health - Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care” Grant (# PNC0000001)

COMPETING INTERESTS

The authors declare no competing interests.

References

1. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance, 1st ed. ed, 2021. . World Health Organization, Geneva.
2. European Parliament. Directorate General for Parliamentary Research Services., 2022. Artificial intelligence in healthcare: applications, risks, and ethical and societal impacts. Publications Office, LU.
3. Ferry, C., Garin, D., Corti, R., Clerc, A., Togni, M., Wenaweser, P., Stortecky, S., Cook, S., 2025. Predicting one-year outcomes after TAVI: evidence from a two-site Swiss registry. BMC Cardiovasc Disord 25, 644. https://doi.org/10.1186/s12872-025-05114-8
4. Khaowroongrueng, V., Kim, T.-E., Park, S.-I., Shin, K.-H., 2025. Application of Real-World Evidence to Support FDA Regulatory Decision Making. AAPS J 27, 98. https://doi.org/10.1208/s12248-025-01082-1
5. Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J.R., Schmid, M.K., Balaskas, K., Topol, E.J., Bachmann, L.M., Keane, P.A., Denniston, A.K., 2019. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 1, e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2
6. Nushi, B., Kamar, E., Horvitz, E., 2018. Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. HCOMP 6, 126–135. https://doi.org/10.1609/hcomp.v6i1.13337
7. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C., 2020. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. Proc ACM Conf Health Inference Learn (2020) 2020, 151–159. https://doi.org/10.1145/3368555.3384468
8. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366, 447–453. https://doi.org/10.1126/science.aax2342
9. On behalf of Topic Group ‘Evaluating diagnostic tests and prediction models’ of the STRATOS initiative, Van Calster, B., McLernon, D.J., Van Smeden, M., Wynants, L., Steyerberg, E.W., 2019. Calibration: the Achilles heel of predictive analytics. BMC Med 17, 230. https://doi.org/10.1186/s12916-019-1466-7

10. Otokiti, A.U., Shih, H.-J., Williams, K.S., 2025. Gender and racial bias unveiled: clinical artificial intelligence (AI) and machine learning (ML) algorithms are fanning the flames of inequity. Oxf Open Digit Health 3, oqaf027. https://doi.org/10.1093/oodh/oqaf027
11. Paulus, J.K., Kent, D.M., 2020. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. npj Digit. Med. 3, 99. https://doi.org/10.1038/s41746-020-0304-9
12. Praz, F., Borger, M.A., Lanz, J., Marin-Cuartas, M., Abreu, A., Adamo, M., Ajmone Marsan, N., Barili, F., Bonaros, N., Cosyns, B., De Paulis, R., Gamra, H., Jahangiri, M., Jeppsson, A., Klautz, R.J.M., Mores, B., Pérez-David, E., Pöss, J., Prendergast, B.D., Rocca, B., Rossello, X., Suzuki, M., Thiele, H., Tribouilloy, C.M., Wojakowski, W., ESC/EACTS Scientific Document Group, Vahanian, A., Mestres, C.-A., Abid, L., Aktaa, S., Akowuah, E.F., Arbelo, E., Asselbergs, F.W., Barbato, E., Boriani, G., Brida, M., Buccheri, S., Byrne, R.A., Chioncel, O., Conradi, L., De Bonis, M., Delgado, V., Franzone, A., Haugaa, K.H., Heidecker, B., Ibanez, B., Iung, B., James, S., Køber, L., Koskinas, K.C., Landmesser, U., Lip, G.Y.H., McEvoy, J.W., Meltzer, G., Messika-Zeitoun, D., Mihaylova, B., Mindham, R., Moelgaard, I., Nielsen, J.C., Owens, G., Pasquet, A.A., Pilgrim, T., Prescott, E., Quintana, E., Rudolph, V., Sadaba, R., Sannino, A., Tanner, F.C., Urena, M., Vaartjes, I., Vrints, C., Wahba, A., Walther, T., Witkowski, A., Zeppenfeld, K., Shuka, N., Kichou, B., Chilingaryan, A.L., Bartko, P.E., Samadov, F., Van De Heyning, C.M., Kusljugic, Z., Kinova, E., Bulum, J., Eftychiou, C., Linkova, H., Fosbøl, E., Bahaa, H., Truusalu, J., Piuhola, J., Donal, E., Petriashvili, S., Rudolph, T.K., Drakopoulou, M., Kertész, A., Guðmundsson, H., Cole, B., Carasso, S., Navazio, A., Sugralimova, M., Bajraktari, G., Kerimkulova, A., Elgdhafi, E.O., Glaveckaite, S., Lebrun, F., Demarco, D.C., Lisii, D.M., Streukens, S., Antova, E., Dalen, H., Kukulski, T., Gavina, C., Popescu, B.A., Bini, R., Ivanov, I., Hudec, M., Bunc, M., Bermejo, T.J., Meurling, C.J.C., Jeger, R., Abid, L., Degertekin, M.M., Nesukay, E.G., Garbi, M., Mullabayeva, G., Grabenwoeger, M., Eynden, F.V., Vojacek, J., Vincentelli, A., Falk, V., Dedeilias, P., Parolari, A., Braun, J., Nikolic, A., Ellensen, V.S., Sousa-Uva, M., Micovic, S., Legarra, J.J., Ferrari, E., Moorjani, N., 2025. 2025 ESC/EACTS Guidelines for the management of valvular heart disease. European Heart Journal 46, 4635–4736. https://doi.org/10.1093/eurheartj/ehaf194
13. Rajkomar, A., Dean, J., Kohane, I., 2019. Machine Learning in Medicine. N Engl J Med 380, 1347–1358. https://doi.org/10.1056/NEJMra1814259
14. Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G.E., Irvine, J., Le, Q., Litsch, K., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S.L., Chou, K., Pearson, M., Madabushi, S., Shah, N.H., Butte, A.J., Howell, M.D., Cui, C., Corrado, G.S., Dean, J., 2018. Scalable and accurate deep learning with electronic health records. npj Digital Med 1, 18. https://doi.org/10.1038/s41746-018-0029-1
15. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y., Ghassemi, M., 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 27, 2176–2182. https://doi.org/10.1038/s41591-021-01595-0

16. Tamburino, C., Barbanti, M., D'Errigo, P., Ranucci, M., Onorati, F., Covello, R.D., Santini, F., Rosato, S., Santoro, G., Fusco, D., Grossi, C., Seccareccia, F., 2015. 1-Year Outcomes After Transfemoral Transcatheter or Surgical Aortic Valve Replacement. Journal of the American College of Cardiology 66, 804–812. https://doi.org/10.1016/j.jacc.2015.06.013
17. Tinetti, M.E., Bogardus, S.T., Agostini, J.V., 2004. Potential pitfalls of disease-specific guidelines for patients with multiple conditions. N Engl J Med 351, 2870–2874. https://doi.org/10.1056/NEJMsb042458
18. Tinetti, M.E., Fried, T., 2004. The end of the disease era. Am J Med 116, 179–185. https://doi.org/10.1016/j.amjmed.2003.09.031
19. Traverso, A., Tiano, D., Corvaglia, A., Dimonte, A., Draetta, E.L., Fabiani, B., Scuri, P., Barbieri, S., Agazzi, M., Arslan, M., Celada, D., Chiabrando, F., Cibrario, L., Cielo, G., Colombo, A., Contini, S., Liberotti, M., Montagna, M., Ogliari, F.R., Palmisano, A., Pisu, F., Serra, D., Varani, D., Vignale, D., Vitali, A.L., Zambello, A., Chiapponi, C., Denti, M., Esposito, A., Tacchetti, C., 2026. Powering responsible artificial intelligence with high-quality real-world data: the S-RACE platform for scalable, multi-specialty clinical research. NPJ Digit Med 9, 6. https://doi.org/10.1038/s41746-025-02132-w
20. Van Genderen, M.E., Kant, I.M.J., Tacchetti, C., Jovinge, S., 2025. Moving Toward Implementation of Responsible Artificial Intelligence in Health Care: The European TRAIN Initiative. JAMA 333, 1483. https://doi.org/10.1001/jama.2025.1335
21. Vickers, A.J., Elkin, E.B., 2006. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 26, 565–574. https://doi.org/10.1177/0272989X06295361
22. Walling, A.M., Asch, S.M., Lorenz, K.A., Roth, C.P., Barry, T., Kahn, K.L., Wenger, N.S., 2010. The quality of care provided to hospitalized patients at the end of life. Arch Intern Med 170, 1057–1063. https://doi.org/10.1001/archinternmed.2010.175
23. Webb, J.G., Blanke, P., Meier, D., Sathananthan, J., Lauck, S., Chatfield, A.G., Jelisejevas, J., Wood, D.A., Akodad, M., 2022. TAVI in 2022: Remaining issues and future direction. Archives of Cardiovascular Diseases 115, 235–242. https://doi.org/10.1016/j.acvd.2022.04.001

Figures

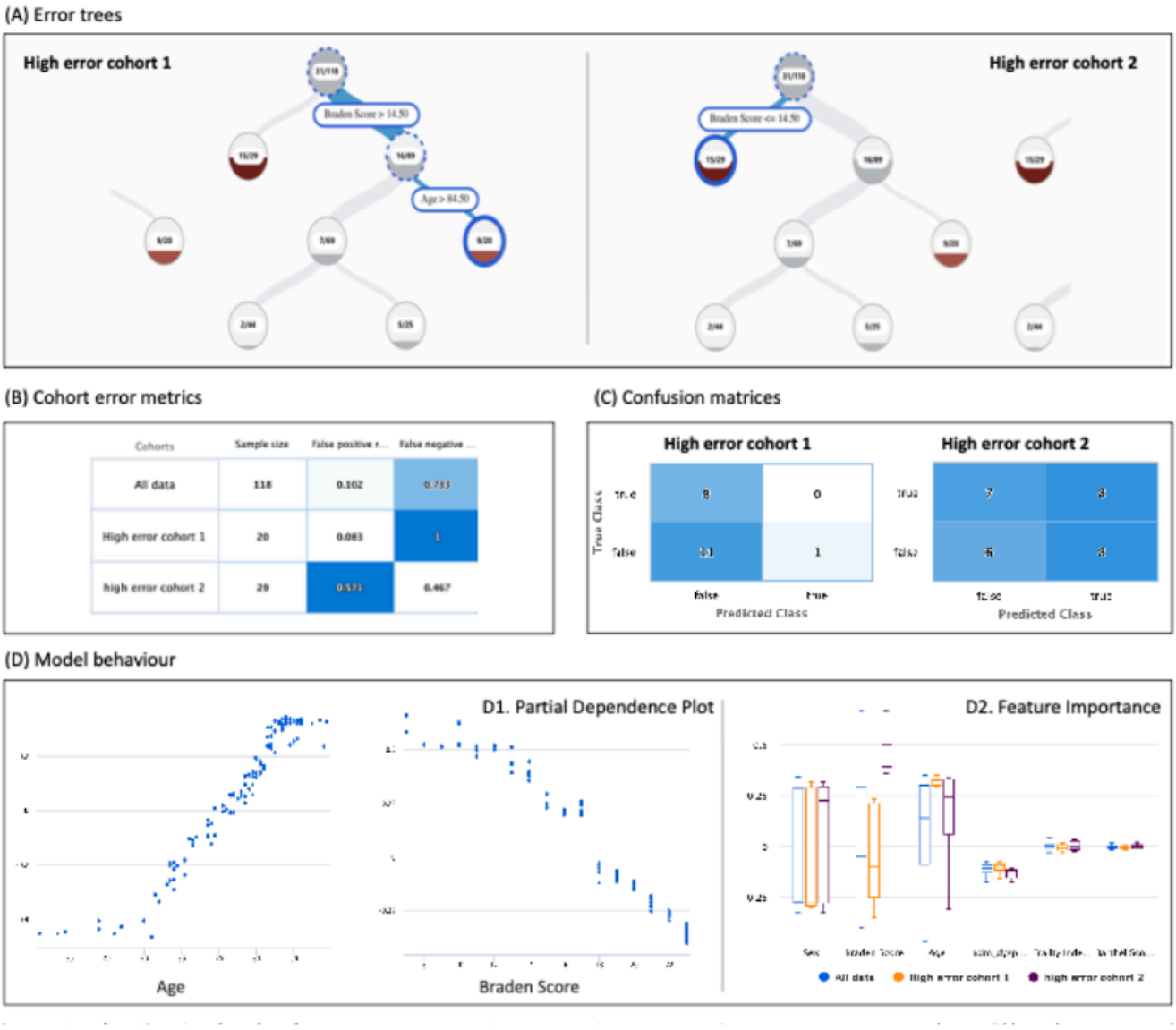


Figure 1

Survival-aware auditing of Med-CLI at a 90-day horizon. (A) Error trees for the labelable evaluation cohort (n = 118) highlighting two clinically interpretable high-error cohorts relative to the global error burden (31/118): patients with Braden Score ≤ 14.50 (15/29 misclassifications; high-error cohort 2) and patients with Braden Score > 14.50 and Age > 84.50 (9/20; high-error cohort 1). The adjacent subgroup with Braden Score > 14.50 and Age ≤ 84.50 shows substantially lower error (7/69), illustrating heterogeneous model reliability across clinically defined strata. (B) Cohort-level error metrics comparing the overall evaluation set and the two high-error cohorts, reporting sample size, false positive rate (FPR), false negative rate (FNR), and selection rate at the selected operating point. (C) Confusion matrices for the two high-error cohorts derived from event-by-90-days classification of the survival endpoint, detailing the distribution of misclassifications within each subgroup.

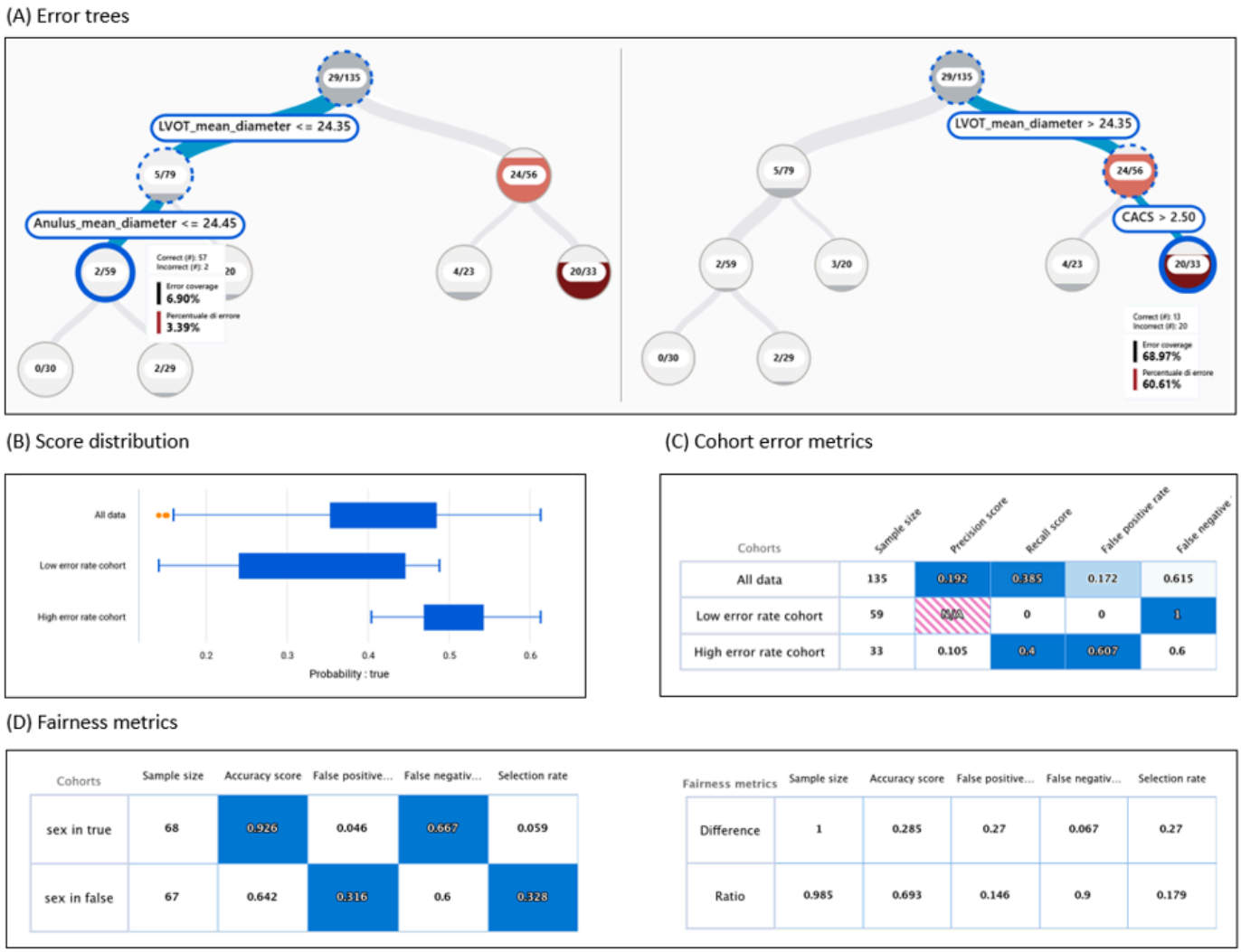


Figure 2

Error analysis and fairness assessment of the TAVI mortality prediction model.

(A) Error analysis trees for the internal temporal validation cohort (n = 135), generated for the eight-feature reduced-CT model. The root node summarizes 29/135 misclassifications. A low-error region defined by *LVOT_mean_diameter*≤ 24.35 and *Annulus_mean_diameter* ≤ 24.45 contains 2 errors among 59 cases (error rate, 3.39%; error coverage, 6.90%). A high-error region defined by *LVOT_mean_diameter* > 24.35 and CACS > 2.50 contains 20 errors among 33 cases (error rate, 60.61%; error coverage, 68.97%).(B) Distribution of predicted probabilities across the overall validation cohort and the error-defined subgroups, showing a rightward shift in the high-error cohort.(C) Cohort-level classification metrics at the selected operating point, reporting sample size, precision, recall, false-positive rate, and false-negative rate for the overall cohort and the two error-defined subgroups. (D) Audit-only sex assessment reporting group-wise performance and derived fairness metrics. Sex was not used by the predictive model and was included only as an example variable for post hoc subgroup

auditing. Observed differences in accuracy, false-positive rate, and selection rate do not establish unfairness, but identify subgroup-specific patterns that warrant further investigation.