scientific reports

Check for updates

OPEN

Global performance of machine learning models to predict all-cause mortality: systematic review and meta-analysis

Felipe Mendes Delpino[1]✉, Ludmila Pereira Pimenta[2], Diego Ferreira Gonzalez[3], Audêncio Victor[2,4], Cinthia Fonseca Araujo[5], Keisyanne De Araujo-Moura[2], J. Jaime Miranda[6], Sandro Rogério Rodrigues Batista[7,8], Alexandre Dias Porto Chiavegatto Filho[2] & Bruno Pereira Nunes[1,9]

We aimed to review the literature on the performance of machine learning models to predict all-cause mortality. The systematic review was protocolled in PROSPERO (CRD42023476567) following PRISMA guidelines. Searches were conducted in PubMed, LILACS, Web of Science, and Scopus databases. Studies predicting all-cause mortality using machine learning were analyzed with random-effects models, with heterogeneity assessed using I^2 statistics and quality evaluated using TRIPOD + AI. The meta-analysis included 88 studies. Most of the studies were from the United States (n = 25) and China (n = 20). Overall pooled AUC was 0.831 (95% CI 0.797–0.865), with extreme heterogeneity (I^2:100%). The majority of the studies included no social variables in the models (89.8%). Subgroup analysis showed similar performance between general population studies and disease-specific populations. Models from high-income countries were similar to those from low- and middle-income countries. Meta-regression showed covariates that affected the results: algorithm, population type, study quality score, and study CI imputation. Equity-oriented sub-group analysis (< 10%) and external validation in other datasets (8.0%) were scarce. Overall, machine learning models showed high performance to predict all-cause mortality, but also highlighted equity gaps. The limitations reduce the potential of public health's evaluation and deployment due to the risk of perpetuation of social disparities. Extreme heterogeneity indicates highly context-dependent performance requiring local validation before implementation assessment.

Keywords Machine learning, Mortality, Prediction, Review, Meta-analysis

Recent technological advancements due to improvements in computing power have allowed the collection of large volumes of data[1,2]. In the health area, these innovations present an opportunity to increase the accuracy of outcome predictions[3,4], such as all-cause mortality, a complex outcome in public health approaches, especially due to the multifactorial nature of this outcome[5–7]. The use of predictive models for all-cause mortality could be useful to provide better care for individuals and populations, mainly to offer approaches to prevent premature mortality and events associated with poor disease management, which could lead to avoidable deaths. Machine learning models can offer new perspectives for predicting all-cause mortality[8–11] considering its ability to deal with complex relationships among variables. However, the performance of these models can be influenced by the context and the representativeness of the available data.

[1]Postgraduate Program in Nursing, Federal University of Pelotas, Gomes Carneiro, 01, Pelotas, Rio Grande do Sul, Brazil. [2]School of Public Health, University of São Paulo, São Paulo, São Paulo, Brazil. [3]Faculty of Medicine, Federal University of Pelotas, Pelotas, Rio Grande do Sul, Brazil. [4]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [5]Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil. [6]Sydney School of Public Health, Faculty of Medicine and Health, University of Sydney, Sydney, NSW 2006, Australia. [7]Faculty of Medicine, Universidade Federal de Goiás, Goiânia, Brazil. [8]Postgraduate Program in Medical Sciences, Faculty of Medicine, University of Brasília, Brasília, Brazil. [9]Department of Health and Kinesiology, University of Illinois Urbana-Champaign, Urbana, IL, USA. ✉email: fmdsocial@outlook.com

Furthermore, the presence of unbalanced classification is a challenging issue in machine learning[12], which occurs when the class distribution is significantly different from 50%, a common situation in epidemiological studies assessing mortality. In particular, predicting mortality in general population samples, which can include both "healthy" individuals and those with diseases, tends to present additional challenges. The variability in the results and the complexity to develop machine learning models for all-cause mortality highlight the need for a comprehensive analysis to better understand the state-of-the-art of these models. Two previous systematic reviews have demonstrated that machine learning has potential for predicting chronic diseases and obesity[13,14] despite the limitations observed.

Systematic reviews and meta-analyses can provide valuable insights for the effectiveness of predictive models from a global health equity perspective. The identification of factors influencing predictive performance across different populations and economic contexts may be particularly valuable to improve the global applicability of these models, as suggested by recent ethical frameworks for AI in health[15]. In the study, we aimed to review the literature on the performance of machine learning models to predict all-cause mortality and to synthesize these results through a meta-analysis.

Methods

We carried out a systematic review and meta-analysis, which was registered in the PROSPERO repository (CRD42023476567), and was conducted following the recommendations of the PRISMA 2020 statement[16].

Search strategy

We searched on October 24, 2023 in the following databases: Pubmed, LILACS, Web of Science, and Scopus. No restrictions were imposed on the year of publication, country, or language of the studies included. If studies were identified in languages other than Portuguese, English, or Spanish, we utilized Google Translate for translation.

Two groups of keywords were used and combined with the Boolean operators 'OR' and 'AND', respecting the specificities of each database. Whenever possible, we used the Medical Subject Heading (Mesh) Major Topic, or, when not available by the database, the searches included results for the articles' titles. We used filters for original articles when available from the databases.

The following keywords were used in the searches: Machine Learning [MeSH] OR Supervised Machine Learning [MeSH] OR Prediction models OR Prediction OR Predictive OR Predict OR Classification OR ML OR Artificial Intelligence [MeSH] OR Natural Language Processing [MeSH] OR Neural Networks, Computer [MeSH] OR Support Vector Machine [MeSH] OR Naive Bayes OR Bayesian learning OR Logistic Models [Mesh] OR Neural network OR Neural networks OR Natural language processing OR Support vector* OR Random forest* OR Boosting OR XGBoost OR Deep learning [Mesh] AND Death [MeSH] OR Mortality [MeSH] OR All-cause mortality. The terms and strategy were adapted according to the specifics of each database, and the complete search strategy is available in the Supplementary Table 1.

Inclusion criteria

We included studies that followed three criteria: predicted all-cause mortality classification as a binary outcome utilizing any machine learning models; involved adults or older adults; and reported the results as Area Under the Curve (AUC) or as true positive, true negative, false positive, and false negative (to calculate the specificity and sensitivity).

Exclusion criteria

We excluded studies that predicted specific mortality (e.g., death by cardiovascular diseases or accidents), studies with children and adolescents, studies with animals, and studies without information to be included in the meta-analysis.

Literature screening and data extraction

The selection of the studies was conducted by two reviewers independently (KAM and AAV), and disagreements were resolved by a third reviewer (FMD). The process began by reading the titles and abstracts of the articles using the Rayyan platform. The second stage involved reading the articles in full and then reviewing the references of the included articles in order to find any new studies. From each included study, we extracted information on the year of publication, location where the study was carried out, sample characteristics, predictor variables, outcome, models used, AUC results, and the best-performing model.

For analytical purposes, studies were classified into two categories based on their sampling approach: general population cohorts: studies recruiting participants from community-based settings regardless of disease status, including population registries, health surveys, or community screening programs); disease-specific cohorts: studies recruiting participants based on specific medical conditions or clinical settings (e.g., patients with heart failure, hospitalized patients, or disease-specific registries).

Quality assessment

To evaluate the individual risk of bias, we utilized an adapted version of the Transparent Reporting of a multivariable prediction model of Individual Prognosis or Diagnosis (TRIPOD + AI)[17]. This updated checklist consists of 27 items that cover various aspects of study reporting. Each item on the TRIPOD + AI checklist is scored, leading to a total score, which can reach 52 points, based on the inclusion of these essential reporting elements. The TRIPOD + AI consists of topics related to title, abstract, introduction, methods, open science, patients and public involvement, results, and discussion. Four reviewers independently conducted the TRIPOD + AI checklist.

Diagnostic criteria

The outcome was all-cause mortality, which was considered when the study evaluated overall mortality rather than specific causes (e.g., cardiovascular mortality). The outcome was chosen because of the need to understand whether machine learning can predict it well, considering that all-cause mortality can occur for different reasons, and to identify which characteristics can influence its occurrence.

Statistical analysis

We conducted a meta-analysis of the Area Under the Curve (AUC) of a Receiver Operating Characteristic (ROC). We chose AUC as our metric because it comprehensively evaluates how well models discriminate between positive outcomes (mortality) and negative outcomes (survival)[18]. The AUC presents values ranging from 0 to 1. A value of 1 signifies a perfect model that can perfectly differentiate between the two classes, while a value of 0.5 indicates performance equivalent to random guessing[18]. AUC values above 0.7 are considered good predictive performance.

We collected each study's AUC value, confidence intervals, and standard errors (SE), based on the model that performed best in the test set or external validation. For studies that reported confidence intervals instead of SE, we calculated it using the following formula: SE = (upper limit − lower limit)/3.92[19]. For studies lacking confidence intervals, we estimated them using the method of Hanley and McNeil[20], which calculates confidence intervals based on AUC values and sample sizes under the assumption of a binomial distribution. We assumed a 95% confidence level and used the relationship between AUC, sample size, and standard error. Sensitivity analyses compared meta-analytic results with and without imputed confidence intervals to assess the impact of this assumption on pooled estimates and heterogeneity measures[20].

In addition to the general meta-analysis, we carried out subgroup analyses: 1- general population vs people with specific diseases or conditions; 2- high-income vs low- and middle-income countries, based on the World Bank classification; 3- sample size: less than 2000 versus 2000 or more; 4- TRIPOD + AI: less than 35 points vs 35 or more; 5- models: Tree-based and Tree Ensemble vs. Neural Networks vs. Linear/Statistical vs Ensemble/Hybrid Models vs other models). Model categories were defined as follows: (1) Tree-based models: individual decision trees and random forests when used as single algorithms; (2) Neural Networks: all artificial neural network architectures including deep learning, convolutional networks, and multilayer perceptrons; (3) Linear/Statistical: logistic regression, Cox regression, and linear discriminant analysis; (4) Ensemble/Hybrid models: combinations of multiple algorithms including XGBoost, LightGBM, and stacked approaches; (5) Other models: single-use algorithms including Disease Severity Models (DSM), Deep Learning System with Multi-head Self-attention Mechanism (DLS-MSM), ICD-based Injury Severity Score (ICISS), Support Vector Machines, and Bayesian Networks. The "Other" category comprised algorithms used by only one study each, limiting statistical power for meaningful comparisons.

The results are presented using the AUC, through a random-effects model, with a 95% confidence interval (95% CI). The heterogeneity of the meta-analysis was assessed using the I^2 statistic, considering values above 75% as high heterogeneity[21,22], which means that the studies are very different from each other, and the combined interpretation of the results should be conducted with caution.

The meta-analyses were conducted using the Python language, via Google Colab, utilizing the Numpy, Pandas, Statsmodels, Matplotlib, and Seaborn libraries. When the same study evaluated different all-cause mortality follow-ups, we considered the longest period for the meta-analysis. Whenever the studies provided such information, we collected the results from the test set or the external validation dataset. The codes created are available at: https://github.com/fmdsocial/reviewsci/blob/main/Revis%C3%A3o_PDJ_Corre%C3%A7%C3%B5es_28_04_2025.ipynb.

To explore sources of heterogeneity, we conducted univariate meta-regression analyses using weighted least squares to examine the relationship between study characteristics and AUC performance. The moderator variables examined included country income level (high-income vs low/middle-income), population type (general vs disease-specific), sample size (≥ 2000 vs. < 2000 participants), TRIPOD + AI quality score (≥ 35 vs. < 35 points), imputation of confidence interval (non-imputed vs imputed), prevalence of the outcome (0–19%, 20–39%, and 40% or more), and machine learning model type (using tree-based models as reference category). Model categories were defined as follows: Tree-based models included individual decision trees, random forests when used as single algorithms; Ensemble models included combinations of multiple algorithms (Random Forest, XGBoost, LightGBM, and hybrid approaches); Neural Networks included all artificial neural network architectures; Linear/Statistical included logistic regression and similar approaches; Other models included single-use algorithms (DSM, DLS-MSM, ICISS, Support Vector Machine, Bayesian Network).

Results

Study selection

From the four databases searched, 33,550 studies were identified, and 17,604 remained after removal of duplicates. Of these, 926 were selected based on inclusion and exclusion criteria by the reviewers. The final step, involving full-text data extraction, resulted in 88 studies being included in the present review (Fig. 1). The main reasons for excluding studies were outcomes other than all-cause mortality, conference abstracts, and articles not reporting AUC metrics.

Studies characteristics

The global distribution of studies is shown in Fig. 2. The United States leads with 25 articles, followed by China (n = 20), Sweden (n = 7), and Taiwan (n = 6). The United Kingdom, Italy, Japan, the Netherlands, and South Korea had between three and five articles, while Finland and Spain had two publications each. Some studies combined databases from different countries (n = 7). In total, 51 studies (58%) were carried out in high-income countries.

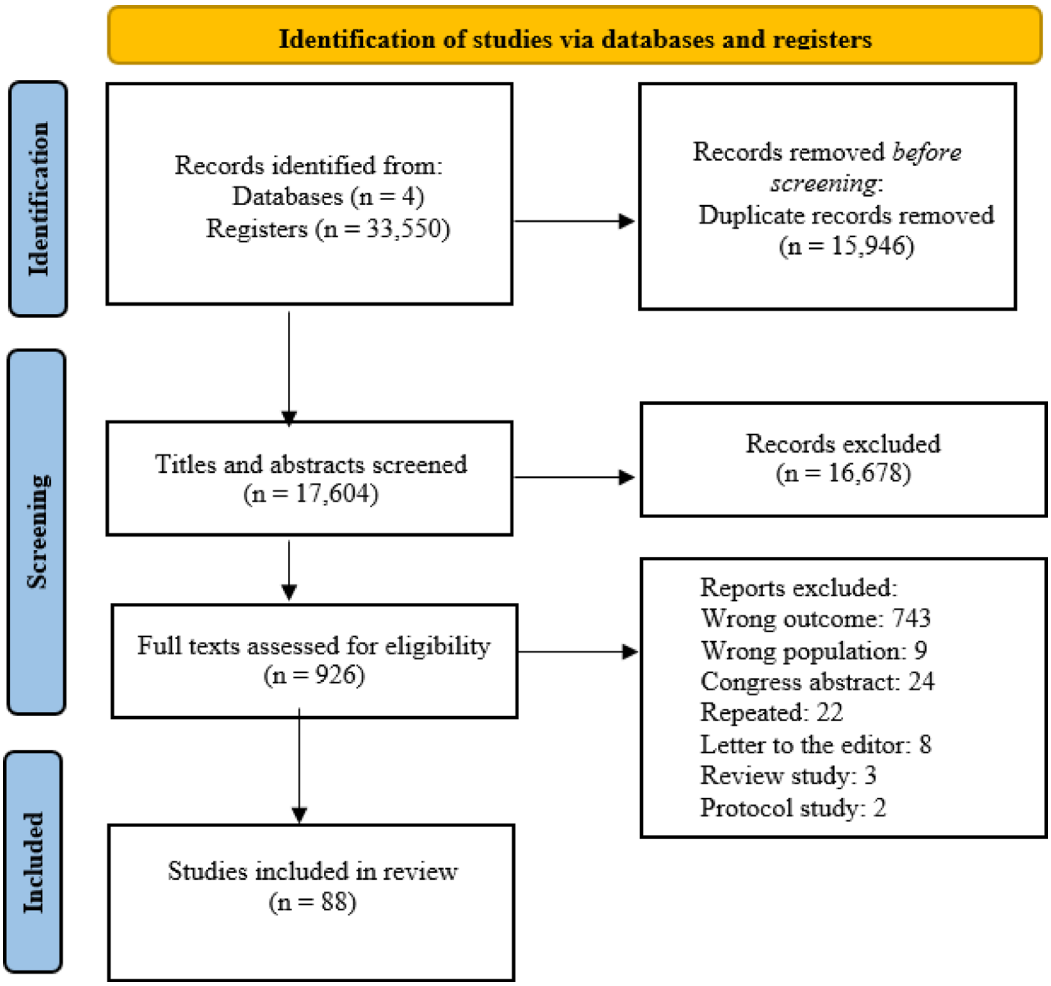

Fig. 1. PRISMA flow diagram of study selection process. Flow diagram showing the identification, screening, eligibility assessment, and inclusion of studies in the systematic review and meta-analysis, following PRISMA 2020 guidelines.

The studies included in the systematic review showed considerable variation in sample size, ranging from 148 to 1,264,000 participants (Table 1). Supplementary Tables 2 and 3 detail the studies included according to the type of outcome. The study with the smallest number of participants was conducted in Sweden[23], with 148 patients discharged from an emergency department. The study with the largest number of participants was conducted in the United States, involving 1,264,000 participants from a synthetic dataset[24]. The average number of participants was approximately 55,840, while the median was 8367. Of the 88 studies, 17 included samples from the general population (19%), with participants with and without disease, and the rest included databases with participants with specific diseases or health conditions.

Qualitative synthesis of predictor variables

Across the 88 studies included in this systematic review, a diverse range of predictor variables was evaluated to predict all-cause mortality, categorized into demographic, clinical, laboratory, imaging, and socioeconomic/behavioral groups. Demographic variables, such as age and gender/sex, appeared in 97.73% (86/88) of studies. Clinical characteristics, including comorbidities (e.g., diabetes, hypertension, cardiovascular disease), vital signs, and medical history, were also highly prevalent, used in 88.64% (78/88) of studies. These variables were prevalent in disease-specific cohorts, where they contributed to high predictive performance, as seen in Díez-Sanmartín et al.[25], which achieved an AUC of 0.99 using XGBoost for kidney transplant patients.

Laboratory biomarkers, such as glucose, cholesterol, troponin, and NT-proBNP, were included in 47.73% (42/88) of studies, with a higher prevalence in studies of chronic diseases (e.g., Takahama et al., 2023, AUC 0.87 using LightGBM for heart failure patients[8]). Imaging data, including ECGs, echocardiograms, and chest radiographs, were less common, appearing in 20.45% (18/88) of studies, predominantly in high-income countries (HICs) with advanced diagnostic infrastructure (e.g., Siegersma et al.[26], AUC 0.96 using deep neural networks with ECGs). Socioeconomic and behavioral factors, such as education, smoking, alcohol consumption, and social support, were the least frequently used, appearing in 27.27% (24/88) of studies. Studies achieving higher AUCs (>0.90) often integrated multiple variables, particularly clinical and laboratory data.

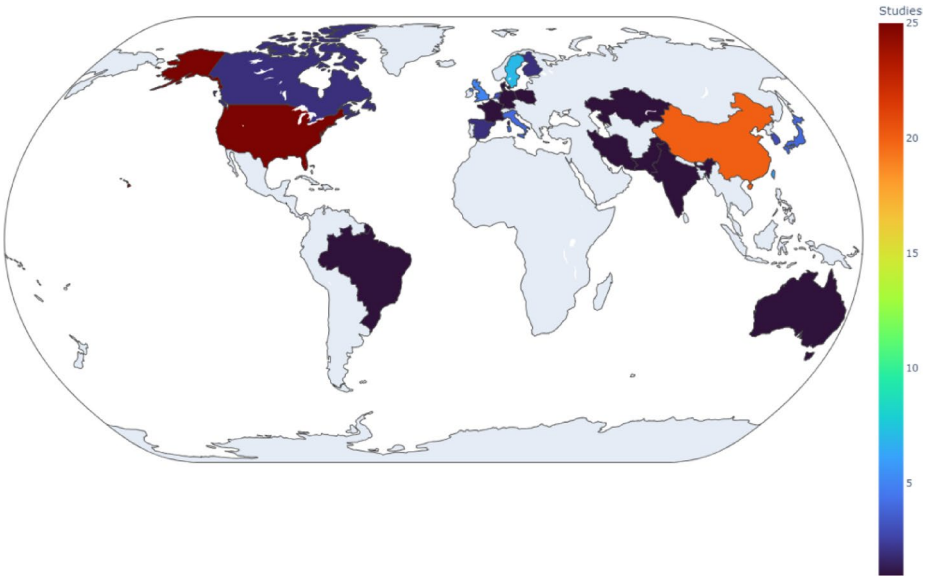

Fig. 2. Global distribution of included studies. World map showing the geographical distribution of the 88 studies included in the systematic review, with the United States (n = 25) and China (n = 20) contributing the most studies.

Algorithms

The studies included in the systematic review utilized various machine learning algorithms to predict mortality and other clinical outcomes. Logistic regression, random forest, and artificial neural networks (ANN) were the most frequently algorithms used in the models. Less frequent models have included algorithms such as Naïve Bayes, Support Vector Machine (SVM), and K-nearest neighbors (KNN). Other algorithms, such as Gradient Boosting Machine (GBM), XGBoost, and LightGBM, also appeared frequently among the studies.

Model predictions

The AUC values ranged from 0.512 to 0.99, five studies (6%) reported an AUC < 0.70. Among the studies with higher AUC, the study by Díez-Sanmartín et al. (2023) using the XGBoost model to predict mortality in patients on the waiting list for kidney transplantation reported an AUC of 0.99[25]. Another study, which used logistic regression to predict mortality among patients hospitalized with diabetes and hypertension, obtained an AUC of 0.97[27]. Some studies showed lower AUCs among patients with diseases. For example, a study on patients with heart failure used a deep learning system based on a multiple self-attention mechanism and obtained an AUC of 0.75 for predicting 365-day mortality[28].

Studies that used datasets from the general population showed, in general, lower AUC values. A study involving a sample of more than 1 million participants showed an AUC of 0.747[24], based on a variety of clinical characteristics as predictors. Another study conducted with 2,291 healthy older adults aged ≥ 70 achieved an AUC of 0.512[29], indicating poor capability to predict all-cause mortality.

Meta-analysis

Figures 3 and 4 show meta-analysis results. The AUC obtained for all-cause mortality prediction was 0.831, 95% CI 0.797 to 0.865, and a heterogeneity of 100%. The AUC was 0.824, 95% CI 0.729 to 0.920, and heterogeneity of 100% in analysis from the general population, regardless of disease status (Figs. 5 and 6). Figures 7 and 8 shows that the AUC was 0.833, 95% CI 0.813 to 0.854, with 99.7% heterogeneity, among the studies carried out with participants with diseases or some health conditions. The results were also similar between high-income countries (0.831, 95% CI 0.788 to 0.874, $I^2 = 100\%$) and low- and middle-income countries (0.830, 95% CI 0.797 to 0.864, $I^2 = 98.0\%$) (Fig. 9). The Fig. 10 shows pooled AUC values for all-cause mortality prediction (Table 2).

Subgroup analysis

Studies with smaller sample sizes (< 2000) showed a slightly higher pooled AUC (0.835, 95% CI 0.799–0.871) compared to larger studies (≥ 2000) (0.830, 95% CI 0.789–0.870), though both demonstrated extreme inter-study variability (I^2 of 96.1% and 100% respectively), see Table 3. Regarding TRIPOD + AI scores, models with lower methodological quality (< 35 points) performed marginally better (AUC 0.838; 95% CI 0.817–0.858) than those with higher quality (≥ 35 points) (AUC 0.814; 95% CI 0.704–0.924), both with significant heterogeneity. Among model types, Linear/Statistical models demonstrated the highest performance (AUC 0.853; 95% CI 0.759–0.947), followed by Ensemble/Hybrid models (AUC 0.829; 95% CI 0.781–0.878), while Tree-based models (AUC 0.832; 95% CI 0.774–0.890), Neural Networks (AUC 0.823; 95% CI 0.793–0.854), and Other models (AUC 0.821; 95% CI 0.734–0.907) also showed strong but slightly lower discriminative ability. Regarding confidence interval imputation, studies with imputed confidence intervals showed higher performance (AUC 0.856; 95%

Author, year and title	Location	Sample characteristics	Predictor variables	Outcome	Models	AUC results and best performing model
Banerjee et al., 2021[49]	United Kingdom	1706 patients with schizophrenia, average age not specified	Antidepressants, second-generation antipsychotics, alcohol/substance abuse, delirium, dementia, cardiovascular disease, diabetes, other physical and mental health issues, social factors like family support	All-cause mortality in schizophrenia patients	Logistic regression, random forests, deep learning models	0.80 Random forest model
Meredith et al., 2002[50]	United States	76,871 incidents, 72,827 survivors, 4044 deceased	Scores based on AIS and ICD-9, including ISS, NISS, APS, maxAIS and their mapped ICD to AIS versions, and the ICD-9 based ICISS score	All-cause mortality	Scoring algorithms based on AIS and ICD, logistic regression used for calibration	0.89 ICISS
Li et al., 2023[28]	China	10,311 patients with heart failure	66 predictors including demographics, pre-existing conditions, treatments received, and other clinical and laboratory details	All-cause mortality within 30 days, 180 days, 365 days, and after 365 days	Deep Learning System based on Multi-head Self-attention Mechanism (DLS-MSM)	0.75 DLS-MSM
Barsasella et al., 2022[27]	Taiwan	58,618 patients, including 25,868 with T2DM, 32,750 with HTN, and 6,419 with both conditions, average age 75.12 ± 13.65 years	67 predictor variables including hospital cost, vital signs and symptoms, comorbidities, demographic characteristics	All-cause mortality	Logistic Regression, Ridge Classifier, Random Forest, K-Neighbors Classifier, Bagging Classifier, Gradient Boosting Classifier	0.97 Logistic Regression
Wang Y. et al., 2021[51]	China	1,200 patients with ESRD undergoing HD, using 36 continuous HD sessions to predict 90-day mortality	64 variables related to hemodialysis sessions, including patient blood pressure recorded multiple times per session	Mortality within 90/180/365 days	LSTM Autoencoder, Logistic Regression, Support Vector Machine, Random Forest, LSTM Classifier, Isolation Forest, Stacked Autoencoder	0.73 LSTM
Díez-Sanmartín et al., 2023[25]	Spain	44,663 adults on kidney transplant waiting list, dialysis < 15 years	Sociodemographic factors (e.g., age, gender, etc.)	Mortality on kidney transplant waiting list	XGBoost, K-Means, Agglomerative Clustering	0.99 XGBoost
Xiong J et al., 2023[10]	China	579 tumor samples from endometrial cancer patients	PCD-related genes	All-cause mortality	LASSO	0.92 LASSO
Lin et al., 2019[52]	Taiwan	48,153 ESRD patients aged ≥ 65 years	Age, sex, urbanization level, occupation, comorbidities (including diabetes, hypertension, etc.)	One-year all-cause mortality	Random Forest, Artificial Neural Networks	0.68 Artificial Neural Networks
Liu et al., 2022[53]	Taiwan, Japan	28,745 patients, 20–60 years old	ECG recordings, AI-predicted LVH	All-cause mortality	Deep learning	0.74 Deep learning
Shi et al., 2012[54]	Taiwan	22,926 patients undergoing surgery for hepatocellular carcinoma	Initial clinical data, surgeon volume, hospital volume	5-year mortality	Artificial Neural Networks and Logistic Regression	0.89 Artificial Neural Networks
L. Shi, X.C. Wang, Y.S. Wang, 2013[55]	China	2,150 elderly patients with intertrochanteric fractures, mean age 81.6 years	Age, gender, nursing home, New Mobility Score, dementia or cognitive impairment, diabetes, cancer, cardiac disease	1-year mortality	Artificial Neural Networks and Logistic Regression	0.87 Artificial Neural Networks
Puddu and Menotti, 2012[56]	Italy	1591 men aged 40–59 years, enrolled in 1960 from rural communities in Crevalcore and Montegiorgio with cardiovascular disease	Age, father life status, mother life status, family history of CVD, job-related physical activity, smoking, BMI, arm circumference, blood pressure, heart rate, forced expiratory volume, serum cholesterol, corneal arcus, diagnoses of CVD, cancer, diabetes, minor ECG abnormalities	45-year all-cause mortality	Cox proportional hazards models, Multilayer Perceptron, AND Neural Networks	0.842 Neural Networks
Harris et al., 2019[57]	United States	107,792 patients undergoing nonemergency primary total hip arthroplasties (THAs) and total knee arthroplasties (TKAs) from the ACS-NSQIP dataset 2013–2014	Demographic and clinical variables such as American Society of Anesthesiologists classification, comorbidities, age, and gender	30-day mortality	LASSO Regression	0.73 LASSO
Takahama et al., 2023[8]	Japan	987 heart failure patients, hospitalized, data from 2013 to 2016	15 variables including troponin levels, blood pressure, BMI, hematocrit, BNP levels, CRP, LDL cholesterol, white blood cell count, diastolic blood pressure, creatinine levels, BUN, LVEF	One-year mortality in heart failure patients	Light Gradient Boosting Machine (LightGBM)	0.87 LightGBM
Arostegui et al., 2018[9]	Spain	1,945 patients with colon cancer who had surgery	Residual tumor, ASA Physical Status, pathologic tumor staging, Charlson Comorbidity Index, intraoperative complications, adjuvant chemotherapy, tumor recurrence	1-year mortality post-surgery	Random Forest, Genetic Algorithms, Classification and Regression Trees (CART)	0.90 Multimodels
Jing et al., 2022[58]	United States	124,360 veterans aged > 50; 93.9% male; mean age 68.2 years	924 predictors including demographics, vital signs, medication classes, disease diagnoses, laboratory results, healthcare utilization	10-year all-cause mortality	Gradient Boosting, Random Forest, Neural Networks, SuperLearner ensemble, LASSO	0.84 Gradient Boosting
Tedesco et al., 2021[29]	Sweden	2,291 healthy older adults aged 70	Anthropometric variables, physical and lab exams, questionnaires, lifestyle, wearable data	All-cause mortality	Logistic Regression, Decision Tree, Random Forest, AdaBoost	0.51 AdaBoost

Continued

Author, year and title	Location	Sample characteristics	Predictor variables	Outcome	Models	AUC results and best performing model
Sakr et al., 2017[59]	United States	34,212 patients, age 54 ± 13 years, 55% male	Fitness data from exercise treadmill stress testing, demographic and clinical variables	All-cause mortality	Decision Tree, Support Vector Machine, Artificial Neural Networks, Naïve Bayesian Classifier, Bayesian Network, K-Nearest Neighbor, Random Forest	0.97 Random Forest
Jones et al., 2021[60]	United States	297,498 encounters, median age 68 years, 95% male, including demographics, vital signs, and 21 laboratory values	Demographic characteristics, 38 comorbid conditions, five vital signs, 21 laboratory values, utilizing PSI variables and excluding mental status and chest imaging	30-day all-cause mortality	Logistic regression, spline models, Extreme Gradient Boosting (XGBoost) incorporating various subsets of predictor variables	0.88 XGBoost
Singh et al., 2022[61]	United States	4,735 patients referred for PET between 2010–2018, median follow-up of 4.15 years	Polar maps of stress and rest perfusion, myocardial blood flow, myocardial flow reserve, spill-over fraction, cardiac volumes, singular indices, sex	All-cause mortality	Deep Learning	0.82 Deep Learning
Lu et al., 2019[62]	United States	PLCO: 10,464 participants, mean age 62.4; NLST: 5,493 participants, mean age 61.7; both included smokers and nonsmokers aged 55–74 years	Chest radiographs	All-cause mortality	Convolutional Neural Network	0.75 and 0.68 Convolutional Neural Network
Siegersma et al., 2022[26]	Netherlands	1,136,113 ECGs from 249,262 individuals, ages 18–85, from UMC Utrecht	12-lead resting ECGs	All-cause mortality	Deep Neural Network	0.96 Deep Neural Network
Ulloa Cerna et al., 2021[63]	United States	34,362 individuals, 812,278 echocardiographic videos	Raw pixel data from echocardiographic videos	All-cause mortality	Convolutional Neural Network	0.84 Convolutional Neural Network
Wang et al., 2021[64]	United States	Patients aged 50 + from a large healthcare system with documented cognitive decline	Clinical notes from EHRs	All-cause mortality	Deep Learning	0.94 Deep Learning
Mohammad et al., 2022[65]	Sweden	139,288 patients admitted for myocardial infarction from the SWEDEHEART registry	Demographic info, medical history, hospital characteristics, lab results	1-year all-cause mortality	Artificial Neural Network	0.85 Artificial Neural Network
Valsaraj et al., 2023[66]	India and Canada	6,083 echos from Taiwan; 997 echos from Alberta, Canada	Echocardiographic data, demographic and clinical data	All-cause mortality (1-year, 3-year, 5-year)	ResNet, CatBoost	0.92 CatBoost
Li et al., 2023[67]	China	Two cohorts: CHNS (8,355 adults > 18 years) and CHARLS (12,711 adults > 45 years)	159 variables from demographics, family, community, socioeconomic status, lifestyle, health conditions, etc	All-cause mortality	Cox Regression, LASSO, Survival Tree, Random Survival Forest, Conditional Inference Forest, glmBoost, Gradient Boosting	0.86 Random Survival Forest
Zhou et al., 2022[68]	Hong Kong	2,560 patients with pulmonary hypertension; median age 63.4 years	Age, average readmission interval, cumulative hospital stay, anti-hypertensive drugs, total bilirubin	All-cause mortality	Random Survival Forest	0.95 Random Survival Forest
Giang et al., 2021[69]	Sweden	71,941 patients with congenital heart disease, mean follow-up time of 16.47 years	Congenital heart disease classifications, comorbidities	All-cause mortality	Neural Networks and Logistic Regression	0.92 Neural Networks
Forte et al., 2021[70]	Netherlands	8,241 patients undergoing valve or CABG operations, mean follow-up of 5 years	Peri-operative clinical parameters	5-year mortality	Super Learner, GLM, XGBoost	0.81 Super Learner
Bergquist et al., 2021[24]	United States	1,264,000 patients	Various clinical predictors	All-cause mortality	Boosted methods, logistic regression, neural networks	0.95 LightGBM
Hernesniemi et al., 2019[71]	Finland	9,066 patients with acute coronary syndrome	Extensive clinical data, GRACE score	Six-month mortality	Logistic regression and XGBoost	0.89 XGBoost
Heyman et al., 2021[23]	Sweden	148 patients discharged from ED, ages 23–106, general ED population	Age, sex, comorbidity score, arrival mode, discharge time, triage priority, imaging during visit	All-cause mortality within 30 days	Logistic Regression, Random Forest, Support Vector Machine	0.95 Logistic Regression
Qiu et al., 2022[72]	United States	47,261 participants, ages across various ranges (data from NHANES 1999–2014)	Demographic, laboratory, examination, questionnaire predictors	All-cause mortality	Gradient Boosted Trees and TreeExplainer	0.92 Gradient Boosting
Niedziela et al., 2021[73]	Poland	17,793 patients treated with primary percutaneous coronary intervention for anterior STEMI	Age, gender, blood pressure, heart rate, Killip class, medical history, in-hospital treatment	6-month all-cause mortality	Logistic Regression, Neural Network	0.81 Neural Network
Mostafaei et al., 2023[74]	Sweden	28,023 dementia-diagnosed patients from the Swedish Registry for Cognitive/Dementia Disorders (SveDem). Median follow-up time was 1053 days for surviving and 1125 days for deceased patients	Age, sex, BMI, MMSE score, dementia type, comorbidities, medications	All-cause mortality	Logistic Regression, Support Vector Machine, Neural Networks	0.74 Support Vector Machine

Continued

Author, year and title	Location	Sample characteristics	Predictor variables	Outcome	Models	AUC results and best performing model
Cui et al., 2022[75]	United States	19,887 lung cancer patients with bone metastases	Age, primary site, histology, race, sex, T stage, N stage, brain metastasis, liver metastasis, cancer-directed surgery, radiation, chemotherapy	3-month mortality	Logistic Regression, XGBoosting Machine, Random Forest, Neural Network, Gradient Boosting Machine, Decision Tree	0.82 Gradient Boosting Machine
Parikh et al., 2019[76]	United States	26,525 adult patients who had outpatient oncology or hematology/oncology encounters at a large academic cancer center and affiliated community practices	Demographic variables, Elixhauser comorbidities, laboratory data	180-day mortality	Logistic Regression, Random Forest, Gradient Boosting	0.88 Random Forest
Tong et al., 2021[77]	China	578 liver cancer patients undergoing RFA	Platelet count (PLT), Alpha-fetoprotein (AFP), age, tumor size, total bilirubin	All-cause mortality	Logistic Regression, DecisionTree, gbm, Gradient Boosting, Forest	0.74 Gradient Boosting model
de Capretz et al., 2023[78]	Sweden	9519 ED chest pain patients, avg. age 59, 47.3% female	Age, sex, ECG, hs-cTnT, glucose, creatinine, hemoglobin	AMI or all-cause death within 30 days	Convolutional neural network, ANN, Logistic Regression	0.94 Convolutional neural network
Tian et al., 2023[79]	China	424 patients with HFmrEF, median follow-up of 1008 days	Age, NYHA class, LVEF, eGFR, NT-proBNP, various lab results	All-cause mortality	XGBoost, Random Forest, SVM	0.92 XGBoost
Mamprin et al., 2021[80]	Netherlands	1,931 TAVI procedures (1,300 at AMC and 631 at CZE)	Clinical predictors including age, health history, and procedural data	One-year mortality	Logistic Regression, Random Forest, CatBoost	0.68 CatBoost
Motwani et al., 2016[81]	Multiple countries	10,030 patients with suspected CAD, 5-year follow-up	25 clinical and 44 CCTA parameters	All-cause mortality within 5 years	Boosted ensemble	0.79 Boosted ensemble
Santos et al., 2019[82]	Brazil	2,808 elderly participants, mean age not specified, general population	37 demographic, socioeconomic, and health profile variables	All-cause mortality within five years	Logistic regression, neural networks, gradient boosted trees, random forest	0.80 Neural Network
Feng et al., 2023[83]	China	8,943 participants, mean age 61.1 years, 79.6% male, all with three-vessel coronary artery disease	18 selected predictors including demographics, medical history, blood tests, and cardiac function assessments	All-cause mortality over 4 years	Random forest	0.81 Random Forest
Yu et al., 2022[84]	China	7,368 patients, age older than 18, post-cardiac surgery including CABG, valvular operations	25 selected predictors including demographics, comorbidities, vital signs, and lab results	4-year all-cause mortality	Logistic regression, neural networks, naïve bayes, gradient boosting, adapting boosting, random forest, bagged trees, extreme gradient boosting	0.80 Adapting boosting
Tamminen et al., 2021[85]	Finland	2,853 unselected prehospital patients encountered in June 2015	NEWS parameters, blood glucose	30-day mortality	Random Forest	0.76 Random Forest
Xu et al., 2023[86]	United States	630 patients with advanced cancer, mean age 59.1 years, 56.19% female	Demographics, clinical data, patient-reported outcomes (PROs) from the Edmonton Symptom Assessment System (ESAS)	180-day mortality	GLM with elastic net, XGBoost, SVM, Neural Network	0.69 XGBoost
Katsiferis et al., 2023[87]	Denmark	48,944 Danish citizens aged 65 and older, spousal bereavement	Age, sex, healthcare expenditures, sociodemographics	All-cause mortality	XGBoost	0.81 XGBoost
Kanda et al., 2022[88]	Japan	24,949 adults with hyperkalemia, multifactorial conditions, aged ≥ 18	Clinical data, lab results, prescription history	All-cause mortality	XGBoost and Logistic Regression	0.82 XGBoost
Lu et al., 2021[89]	Australia	68,889 patients, mean age 76 years, 54% female	Age, sex, medication history, disease history	All-cause mortality	Gradient Boosting Machine, Multi-Layer Neural Network, Support Vector Machine	0.75 Gradient Boosting
Scrutinio et al., 2020[90]	Italy	1,207 patients with severe stroke, average follow-up 988 days, 15.7% 3-year mortality rate	Age, comorbidities (e.g., diabetes, CAD), functional measures	3-year mortality	Random Forest, Logistic Regression	0.93 Random Forest
Li et al., 2020[91]	China	1,244 patients, mean age 63.8, 78.4% male, 75.18% received reperfusion therapy	Comprehensive clinical dataset including demographics and treatment details	1-year mortality after anterior STEMI	GaussianNB, Logistic Regression, KNN, Decision Tree, Random Forest, XGBoost	0.94 XGBoost
Guo et al., 2022[92]	China	751 patients diagnosed with spontaneous intracerebral hemorrhage at West China Hospital	Clinical presentations, laboratory data (e.g., monocyte and lymphocyte levels), radiographic data, Glasgow Coma Scale, hematoma volume, location, age	90-day mortality	Logistic Regression, Category Boosting, Support Vector Machine, Random Forest, Extreme Gradient Boosting	0.84 Logistic Regression
Wang et al., 2023[11]	United States	1229 patients from the MIMIC-IV database, including critical pulmonary embolism patients with or without septic or other cardiopulmonary complications	Age, gender, DVT, VTE history, hematocrit, hemoglobin, anion gap, heart rate, blood pressure, respiratory rate, congestive heart failure, hypertension, atrial fibrillation, vasopressor	All-cause mortality within 30 days	Logistic Regression and XGBoost	0.82 XGBoost
Continued						

Author, year and title	Location	Sample characteristics	Predictor variables	Outcome	Models	AUC results and best performing model
Li et al., 2023[93]	China	451 older patients with CAD, IGT, and DM from a hospital cohort, split into a training group (308) and a validation group (143). Median age 86 years. Majority were males	Demographics, comorbidities (e.g., CHF, DM, hypertension), laboratory tests (e.g., glucose, NT-proBNP), and medications used during hospitalization (e.g., statins, beta blockers)	All-cause mortality	Logistic Regression, Gradient Boosting, Random Forest, Decision Tree	0.84 Gradient Boosting
Asrian et al., 2024[94]	United States	3751 hip fracture patients from the MIMIC-IV database. Includes demographics like age and basic lab tests. Average age was 73	Age, glucose, red blood cell distribution width, mean corpuscular hemoglobin concentration, white blood cells, urea nitrogen, prothrombin time, platelet count, calcium levels, and partial thromboplastin time	1-, 5-, and 10-year mortality	LightGBM, Random Forest, Logistic Regression	0.79 LightGBM
Ivanics et al., 2023[95]	Canada, United Kingdom, United States	Adults who underwent primary liver transplants from Jan 2008 to Dec 2018: Canada (n = 1214), UK (n = 5287), US (n = 59,558)	Harmonized pre-transplant variables like recipient BMI, donor age, MELD score, etc., across three national registries	90-day mortality	LASSO, Ridge, ElasticNET, LightGBM	0.74 Ridge
Behnoush et al., 2023[96]	Iran	8,493 hypertensive patients undergoing CABG, mean age 68.27 years, 63.86% male, 46.84% with diabetes, 38.61% with a family history of CAD	Age, total ventilation hours, ejection fraction, hemoglobin, total cholesterol, LDL-C, HDL-C, triglycerides, fasting blood glucose, creatinine, BMI, and several perioperative factors like ICU hours and cardiopulmonary pump use	1-year mortality	Logistic Regression, Extreme Gradient Boosting, Naïve Bayes, Random Forest, Artificial Neural Network	0.82 Logistic Regression
Kampaktsis et al., 2023[97]	United States	1033 recipients (median age 34 years, 61% male) of isolated heart transplants analyzed from the UNOS database, 2000–2020	Variables included recipient, donor, procedural characteristics, post-transplant predictors, selected using SHapley Additive exPlanations (SHAP)	1-year mortality	CatBoost	0.80 CatBoost
Lin et al., 2020[98]	Taiwan	1,903 patients from Wan Fang Hospital and Taipei Medical University Hospital, diagnosed with chronic liver diseases such as cirrhosis and hepatic coma	Variables included clinical scores (e.g., MELD score), laboratory data (e.g., bilirubin, creatinine), and demographic data	All-cause mortality	Random Forest, Adaptive Boosting, other machine learning models	0.85 Random Forest
Lin et al., 2023[99]	United States	63,215 asymptomatic patients from four centers, median age 54, 68% male, median follow-up 12.6 years	Age, sex, race, CAD risk factors, CAC score, CAC density, number of calcified vessels	10-year all-cause mortality	XGBoost	0.82 XGBoost
Liu et al., 2022[100]	China	340 patients with sepsis-induced CRS at Shanghai Tongji Hospital, aged ≥ 18, data from 2015–2020	Age, SOFA score, myoglobin levels, vasopressor use, mechanical ventilation	1-year mortality	Random Forest, Support Vector Machine, Gradient Boosted Decision Tree	0.85 Random Forest
Forssten et al., 2021[101]	Sweden	124,707 traumatic hip fracture cases, aged 18 or older, between 2008 and 2017	Age, sex, ASA classification, CCI, RCRI, various comorbidities, type of fracture, surgical procedure	1-year postoperative mortality	Logistic Regression, Naive Bayes, Support Vector Machine, Random Forest	0.74 Random Forest
Alimbayev et al., 2023[102]	Kazakhstan	472,950 patients diagnosed with diabetes mellitus, collected between 2014–2019 from the Unified National Electronic Health System of Kazakhstan	Age, duration of diabetes, hypertension, sex, and other comorbidities like coronary heart disease and cerebrovascular accident	1-year mortality	Gaussian Naïve Bayes, K-nearest neighbors, Logistic Regression, Random Forest, AdaBoost, Gradient Boosting, XGBoost, Linear Discriminant Analysis, Perceptron	0.80 XGBoost
El-Bouri et al., 2023[103]	United Kingdom	2,183 venous thromboembolism patients; 1235 pulmonary embolism cases; mean age 69 years	Neutrophils, white blood cell counts, C-reactive protein, haemoglobin, age, O2 saturation, heart rate, blood pressure	30-day, 90-day, and 365-day mortality	Random Forest, XGBoost, logistic regression	0.73 Random Forest
Park et al., 2022[104]	South Korea	4,312 patients with acute heart failure, median age 73, 55% male, median left ventricular ejection fraction 38%	19 clinical predictors (e.g., age, sex, blood pressure) and 8 echocardiographic parameters (e.g., LV ejection fraction)	All-cause mortality at 3 years	CoxBoost	0.76 CoxBoost
Penso et al., 2021[105]	Italy	471 patients with severe AS, undergoing TAVI, median age 81	83 pre-TAVI clinical and echocardiographic variables	All-cause mortality at 5 years	Random forest, XGBoost, Multilayer perceptron, Logistic regression	0.79 Multilayer perceptron
Guo et al., 2021[106]	United States	34,575 patients with liver cirrhosis	41 health variables including demographic and laboratory data	All-cause mortality at 90, 180, and 365 days	Deep Neural Networks (DNN), Random Forest (RF), Logistic Regression (LR)	0.86 Random Forest
Abedi et al., 2021[107]	United States	7,144 patients post-stroke	37 predictors including demographics, medical history, laboratory data	All-cause mortality within 1, 3, 6, 12, 18, and 24 months	Logistic Regression, Extreme Gradient Boosting, Random Forest	0.82 Random Forest
Zhou et al., 2023[108]	China	706 patients, median age 66, 57% male, diagnosed with mitral regurgitation	Age, blood pressure, P-wave duration, lab values (e.g., albumin, creatinine), echocardiographic measurements (e.g., LVEF, LADs)	All-cause mortality	Gradient Boosting Machine, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Artificial Neural Networks	0.80 Gradient Boosting Machine

Continued

Author, year and title	Location	Sample characteristics	Predictor variables	Outcome	Models	AUC results and best performing model
Rauf et al., 2023[109]	Pakistan	2,184 patients with mitral stenosis and atrial flutter, 81.85% females, median age 65	Mitral valve area, right ventricular systolic pressure, pulmonary artery pressure, left ventricular ejection fraction, NYHA class, surgery	All-cause mortality	Gradient Boosting Machine, Decision Tree, Support Vector Machine, Random Forest, Artificial Neural Network	0.84 Gradient Boosting Machine
Shi et al., 2022[110]	China, United Kingdom	2,846 patients with acute pancreatitis, median age 46 years	Demographic info, lab tests on admission (WBC, creatinine, etc.), clinical severity scores (APACHE II, SOFA)	Mortality	Random Forest	0.89 Ranfom Forest
Zhou et al., 2021[111]	China	381 heart transplant recipients, average age 43.8 years	Albumin, recipient age, left atrium diameter, red blood cells, hemoglobin, lymphocyte%, smoking history, use of rhBNP, Levosimendan, hypertension, cardiac surgery history, malignancy, endotracheal intubation history	1-year mortality	Logistic Regression, Support Vector Machine, Random Forest, Extreme Gradient Boosting, Adaptive Boosting, Gradient Boosting Machine, Artificial Neural Network	0.80 Random Forest
Lee et al., 2021[112]	South Korea	22,182 AMI patients, mean age 64 years, 71.8% male	Demographics, clinical predictors, lab results, history of heart disease, interventions, medications	1-year all-cause mortality	Decision Trees, Logistic Regressions, Deepnets, Random Forest	0.92 Random Forest
Tran et al., 2023[113]	France	534 stage 4–5 CKD patients, median age 72 years, 55% male	Age, ESA usage, cardiovascular history, smoking status, vitamin D levels, PTH levels, ferritin levels	2-year all-cause mortality	Bayesian Network, Deep Learning, Logistic Regression, Random Forest	0.81 Bayesian Network
Raghunath et al., 2020[114]	United States	253,397 patients with 1,169,662 12-lead resting ECGs from a large regional health system, 99,371 events occurred over a 34-year period	12-lead ECG voltage–time traces, age, sex	1-year all-cause mortality	Deep Neural Network	0.88 Deep Neural Network
Kawano et al., 2022[115]	Japan	116,749 participants from health checkups; age, sex, and other health data collected	Age, sex, smoking, AST levels, alcohol consumption, and other health checkup data	5-year all-cause mortality	Gradient Boosting Decision Tree (XGBoost), Neural Network, Logistic Regression	0.81 XGBoost
Weng et al., 2019[116]	United Kingdom	502,628 participants aged 40–69, recruited from the general population	60 predictor variables including demographics, lifestyle factors, and clinical measures	Premature all-cause mortality	Deep Learning, Random Forest, Cox regression	0.79 Deep Learning
Zhou et al., 2021[117]	China	1,241 patients with end-stage renal disease (ESRD) on peritoneal dialysis; age range 18 + , routine follow-up for 12 + months	Age, sex, chronic heart disease, diabetes, malignancy, systolic and diastolic blood pressure, cholesterol levels, serum albumin, hemoglobin, blood urea nitrogen, serum creatinine	Premature all-cause mortality	Logistic Regression, Classic Artificial Neural Network, Mixed Artificial Neural Network	0.79 Artificial Neural Network
Huang et al., 2017[118]	Taiwan, China	3,632 breast cancer patients from a retrospective cohort study, underwent surgery between 1996 and 2010	Age, Charlson Comorbidity Index (CCI), chemotherapy, radiotherapy, hormone therapy, surgery volumes of hospital and surgeon	5-year mortality after breast cancer surgery	Artificial Neural Network, Multiple Logistic Regression, Cox Regression	0.72 Artificial Neural Network
Sheng et al., 2020[119]	China	5,351 patients in training cohort and 5,828 in testing cohort from 97 renal centers, all new hemodialysis patients	Demographic info, disease diagnoses, comorbidities, and lab results collected at dialysis start and 0–3 months post-start	First-year all-cause mortality	XGBoost, Random Forest, and Logistic Regression	0.85 XGBoost
Zachariah et al., 2022[120]	United States	2,041 patients with advanced cancer, median age 62.6 years	Demographic data, lab results, diagnostic codes, past medical history	3-month mortality	XGBoost	0.81 XGBoost
Unterhuber et al., 2021[121]	Germany, Italy	1,998 patients from LIFE-Heart Study (Germany), 772 from PLIC Study (Italy), aged 41–85, at increased cardiovascular risk	92 plasma proteins measured, alongside clinical risk scores like SCORE and Framingham, and other clinical data	All-cause mortality	XGBoost, Neural Network	0.94 Neural Network
Wu et al., 2024[122]	China	11,894 patients aged ≥ 65 years who underwent non-cardiac surgery across 20 tertiary hospitals	Preoperative risk factors including medical history (stroke, chronic diseases), laboratory data (mononuclear cell ratio, total cholesterol), and preoperative assessments	6-month mortality	Random Forests, Support Vector Machine, Decision Tree, Naive Bayes	0.97 Random Forest
Hwangbo et al., 2022[123]	South Korea	19,435 patients assessed from the IST-1 dataset, with 8,787 included after exclusions for various reasons such as treatment type and missing data	18 variables including age, sex, level of consciousness, underlying conditions, and imaging findings	6-month all-cause mortality	Stacking Ensemble, Extreme Gradient Boosting, KNN, SVM, Naive Bayes, Random Forest, Logistic Regression	0.78 Stacking Ensemble

Continued

Author, year and title	Location	Sample characteristics	Predictor variables	Outcome	Models	AUC results and best performing model
Ross et al., 2016[124]	United States	1,755 patients undergoing elective coronary angiography at Stanford University Medical Center or Mount Sinai Medical Center from 2004 to 2008	Demographics, clinical comorbidities, medications, lab tests, physical exam variables, socioeconomic variables, genomic markers	All-cause mortality	Elastic Net, Random Forest	0.76 Random Forest
Wang et al., 2019[125]	United States	40,711 completed cases from the Cardiovascular Disease Life Risk Pooling Project, which includes various demographics, physiological test results, medication status, socio-behavioral factors, and mortality indicators	Demographics, physiological tests, medication status, socio-behavioral factors	All-cause mortality	Naïve Bayes, Logistic Regression, Support Vector Machine, Random Forest	0.89 Random Forest

Table 1. Characteristics and main results of included studies (n = 88). Detailed overview of study characteristics including author, location, sample size, predictor variables, outcomes, machine learning models used, and AUC results for each of the 88 studies included in the systematic review.

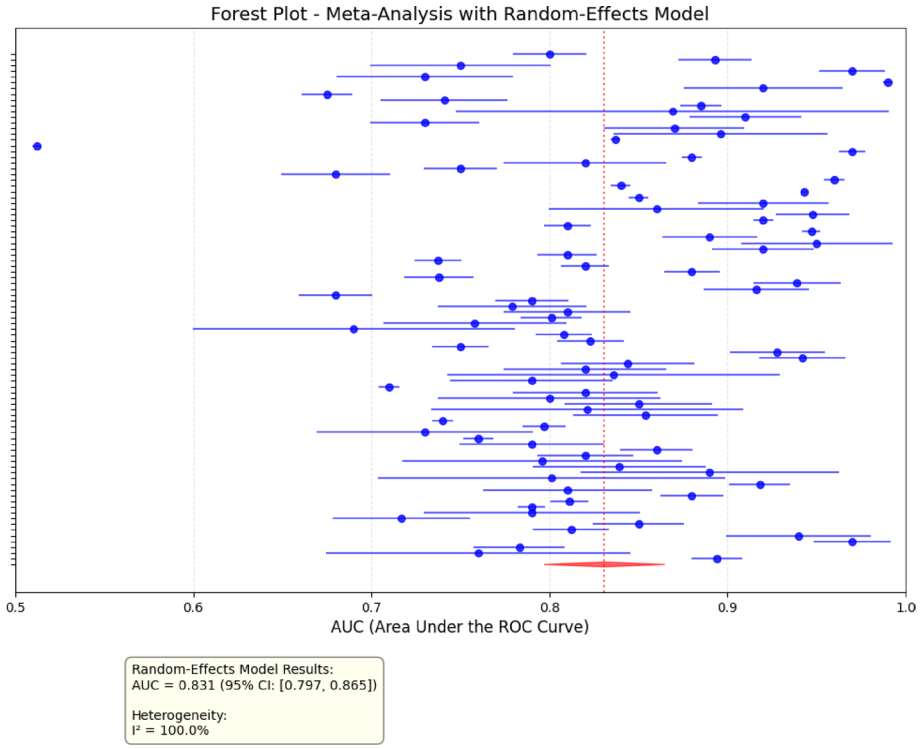

Fig. 3. Forest plot of individual study AUC values and pooled meta-analysis results. Forest plot displaying AUC values for each included study with 95% confidence intervals and the overall pooled estimate using a random-effects model (AUC 0.831, 95% CI 0.797–0.865, I^2: 100%).

CI 0.823–0.890) compared to non-imputed studies (AUC 0.815; 95% CI 0.768–0.862). All model categories and imputation groups showed extreme heterogeneity ($I^2 > 98\%$).

Meta-regression

Meta-regression analysis showed significant study-level covariates associated with AUC performance (Table 4). Studies with disease-specific populations demonstrated significantly higher AUC performance compared to general population studies (reference group) ($\beta = -0.192$, $p < 0.001$). Studies with lower methodological quality scores (< 35 TRIPOD + AI points) showed higher AUC values ($\beta = -0.134$, $p < 0.001$). Neural networks outperformed tree-based models ($\beta = 0.136$, $p < 0.001$). Studies with imputed confidence intervals showed significantly higher AUC values compared to those with non-imputed confidence intervals ($\beta = 0.109$, $p = 0.009$). Regarding mortality prevalence, neither studies with 20–39% mortality rates ($\beta = 0.082$, $p = 0.058$) nor those with ≥ 40% mortality rates ($\beta = -0.025$, $p = 0.732$) showed significant differences in AUC performance compared to studies with 0–19% mortality rates. Country income level, sample size, linear/statistical models, ensemble/hybrid models, and other model types were not significant variables in the model's performance.

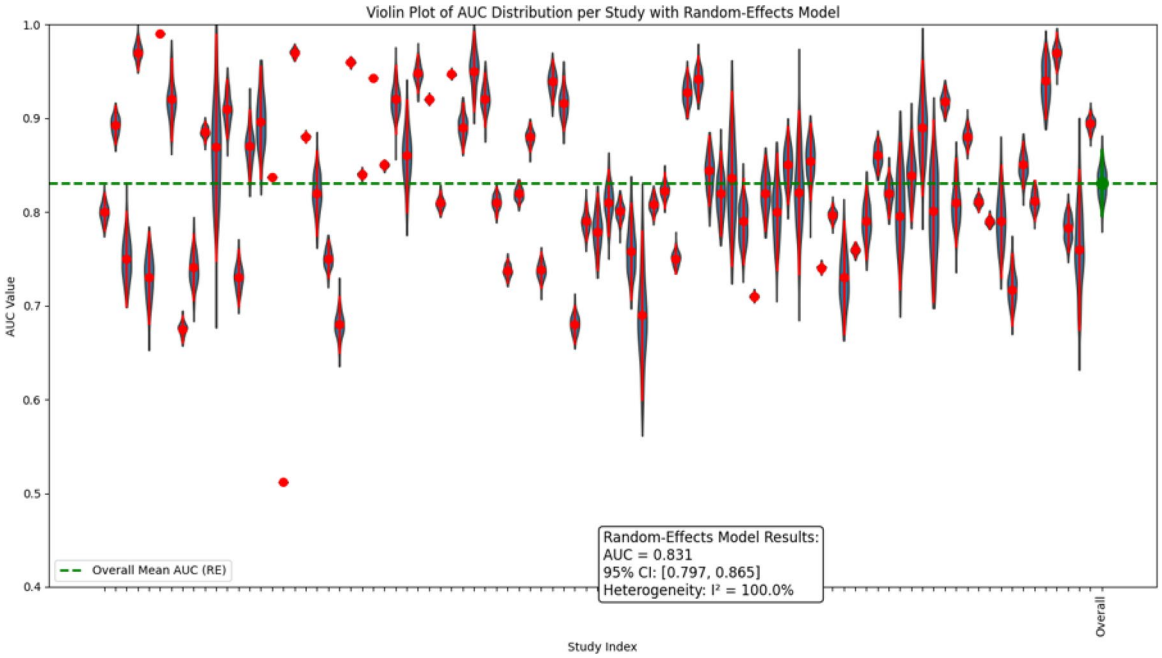

Fig. 4. Distribution of AUC values across all included studies. Violin plot showing the distribution of AUC values from individual studies with the overall meta-analysis result indicated by the central line.

Risk of bias

The TRIPOD + AI scale revealed a diverse range of final scores, ranging from 23 to 45. The lowest score was 23, with the highest reaching 45, while the most frequent score ranged is 31–33, where a significant number of studies cluster, as indicated by the histogram's peak. Specifically, scores of 31 and 33 are notably prevalent, with 18 and 14 studies, respectively, whereas other ranges, such as 23–24, 25–27, and 45, are less common, with only 1–3 studies each (Fig. 11 and Table 2).

Item-level analysis highlighted both strengths and limitations. Studies consistently reported items like Title, Abstract, Background, Objectives, Data, Participants, Outcome, Predictors, Sample Size, Analytical Methods, Ethical Approval, Funding, Conflict of Interest, Model Development, and Interpretation, with 90–100% conformity. However, significant gaps were identified in several areas: only 6.5% of studies addressed Class Imbalance (8c), 9.8% reported Patient and Public Involvement (12f.), 13% discussed Model Updating (18c), 16.3% provided Code Sharing (12e), and 18.5% addressed Data Sharing (12d). Additionally, Fairness (9a, 21.7%), Model Output (9b, 23.9%), Training vs. Evaluation (9c, 26.1%), Protocol (12c, 29.3%), and Usability in Current Care (18f., 30.4%) were frequently underreported.

Equity assessment

Our systematic analysis of equity-related reporting across the 88 included studies revealed substantial gaps in algorithmic fairness considerations (Table 5). The vast majority of studies (89.8%, n = 79) included no social determinants of health variables, with only 10.2% (n = 9) incorporating socioeconomic status, 3.4% (n = 3) including race/ethnicity, and 3.4% (n = 3) reporting education levels. Demographic diversity reporting was limited, with 86.4% (n = 76) of studies providing only basic age and sex data. No studies conducted stratified performance analysis by race/ethnicity or socioeconomic status, and only 2.3% (n = 2) performed sex-stratified analysis. External validation in diverse populations was rare, with 92.0% (n = 81) of studies relying solely on internal validation methods.

Discussion

This systematic review found critical equity gaps suggesting reduced potential of machine learning models to predict all-cause mortality, especially for public health's evaluation and deployment due to the risk of perpetuation of social disparities within these models. Most studies (89.8%) excluded social determinants such as race, education, and income, while none conducted racial/ethnic and socioeconomic sub-group performance analyses. Machine learning models demonstrated high overall performance across diverse populations and economic contexts. However, extreme heterogeneity indicates highly context-dependent results requiring local validation and effectiveness assessment before implementation. Performance was comparable between the general population and disease-specific studies. Nonetheless, this finding has limited generalizability given that only 19% of studies included general population samples.

Our systematic equity assessment showed a lack of algorithmic fairness considerations across the included studies, which represents a fundamental limitation for clinical and public health implementation. The underrepresentation of social determinants variables can reflect multiple barriers, including limited data

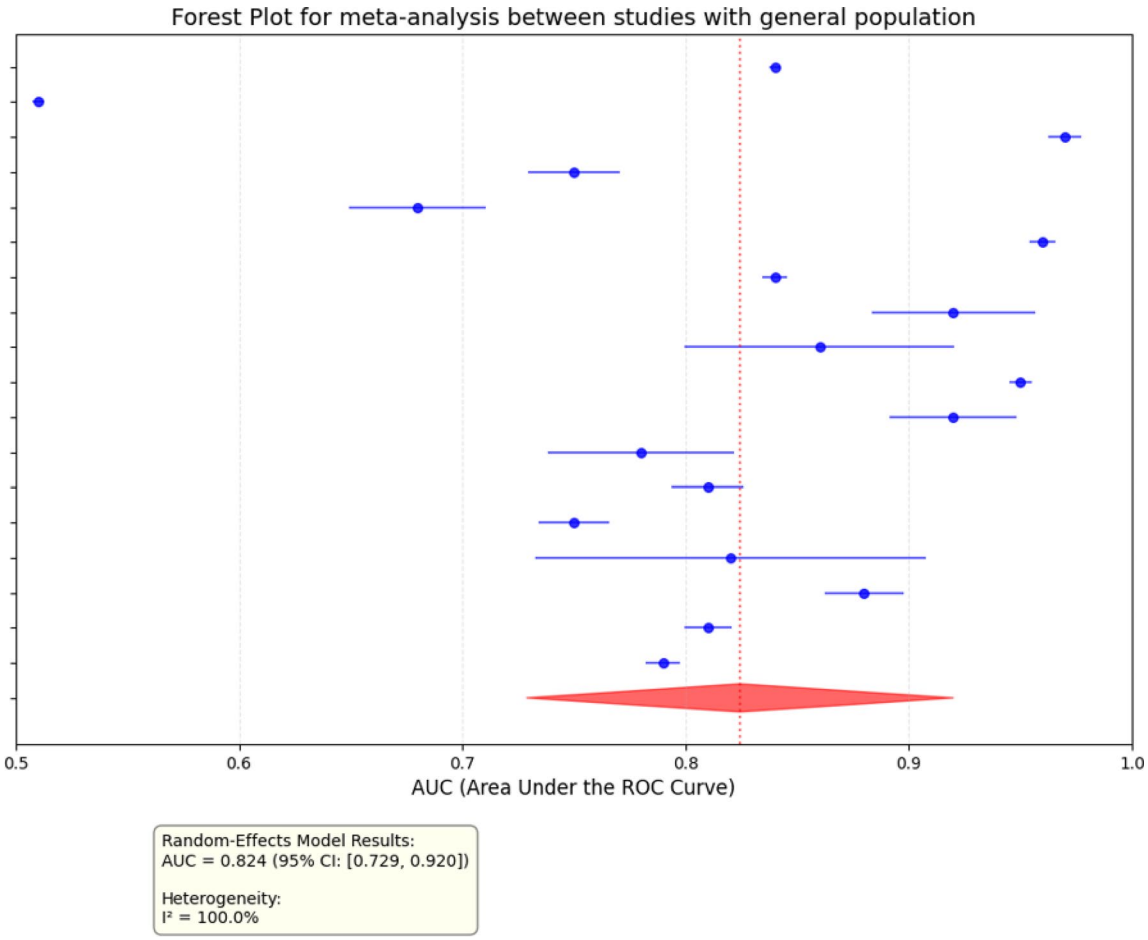

Fig. 5. Forest plot of AUC values for general population studies. Forest plot showing AUC values and 95% confidence intervals for studies conducted in general population cohorts (n = 17) with pooled estimate (AUC 0.824, 95% CI 0.729–0.920, I^2: 100%).

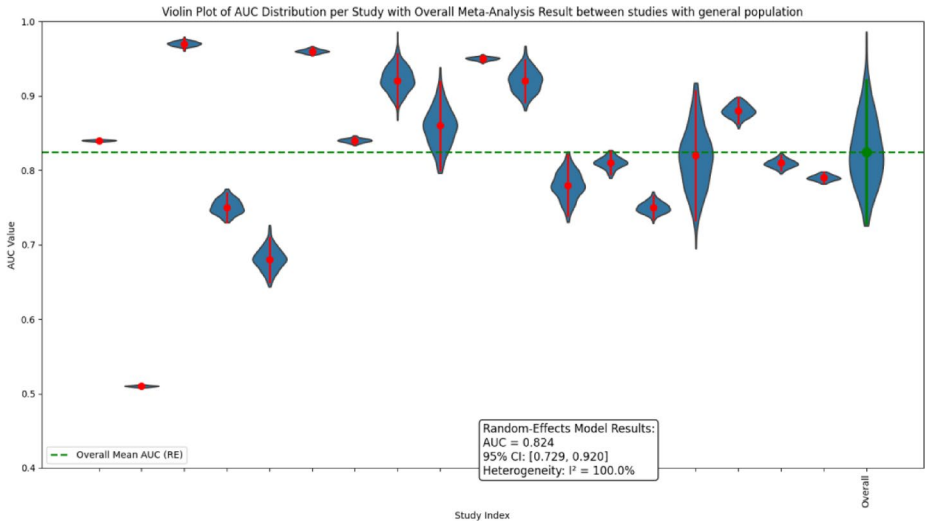

Fig. 6. Distribution of AUC values in general population studies. Violin plot displaying the distribution of AUC values specifically for studies conducted in general population samples with the pooled estimate indicated.

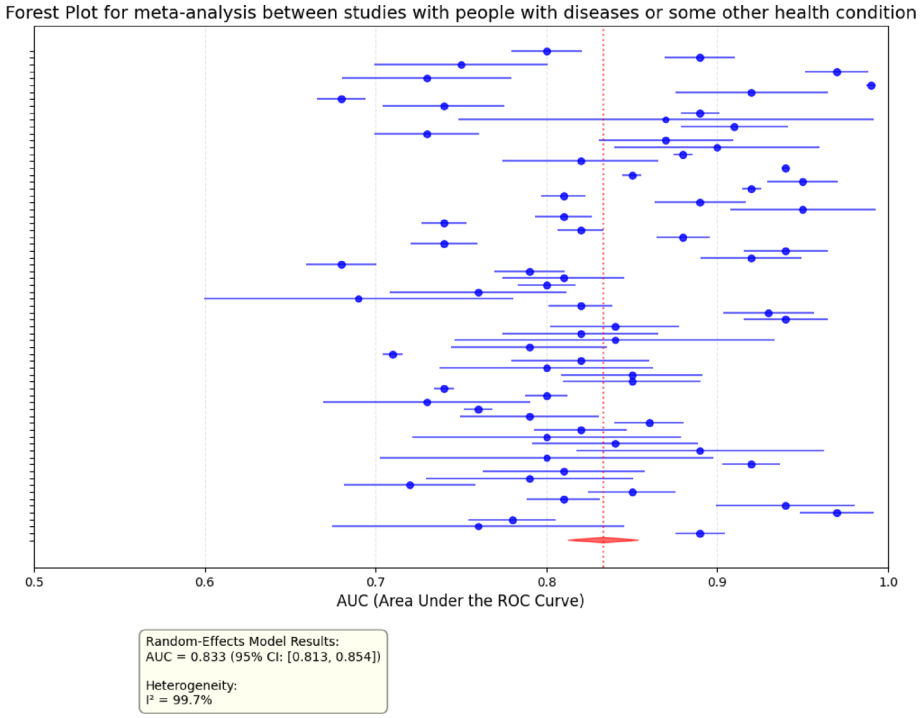

Fig. 7. Forest plot of AUC values for disease-specific population studies. Forest plot showing AUC values and 95% confidence intervals for studies conducted in disease-specific populations (n = 71) with pooled estimate (AUC 0.833, 95% CI 0.813–0.854, I^2: 99.7%).

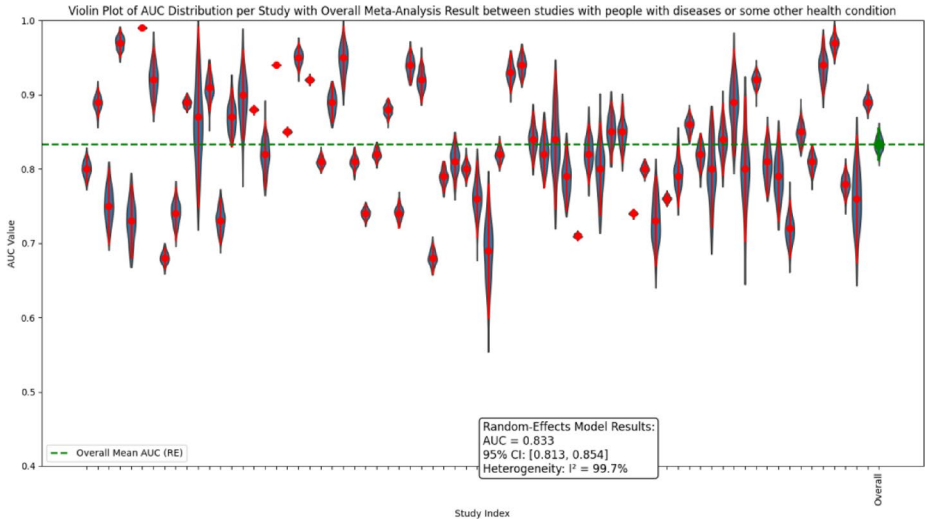

Fig. 8. Distribution of AUC values in disease-specific population studies. Violin plot displaying the distribution of AUC values for studies conducted in disease-specific populations with the pooled estimate indicated.

availability in electronic health records, prioritization of clinical over social variables in model development pipelines, and potential implicit bias in variable selection processes. The absence of race/ethnicity, education, and socioeconomic status as predictors is particularly concerning given their well-established associations with mortality risk and health care accessibility. To address these limitations, we recommend establishing minimum reporting standards for equity variables in future machine learning studies, including mandatory documentation of: (1) availability and inclusion for social determinants variables, (2) demographic representativeness of training datasets, and (3) stratified model performance across relevant demographic and socioeconomic subgroups.

Machine learning models achieved comparable performance in disease-specific and general populations, suggesting broad applicability through different mechanisms. While disease-specific populations offer

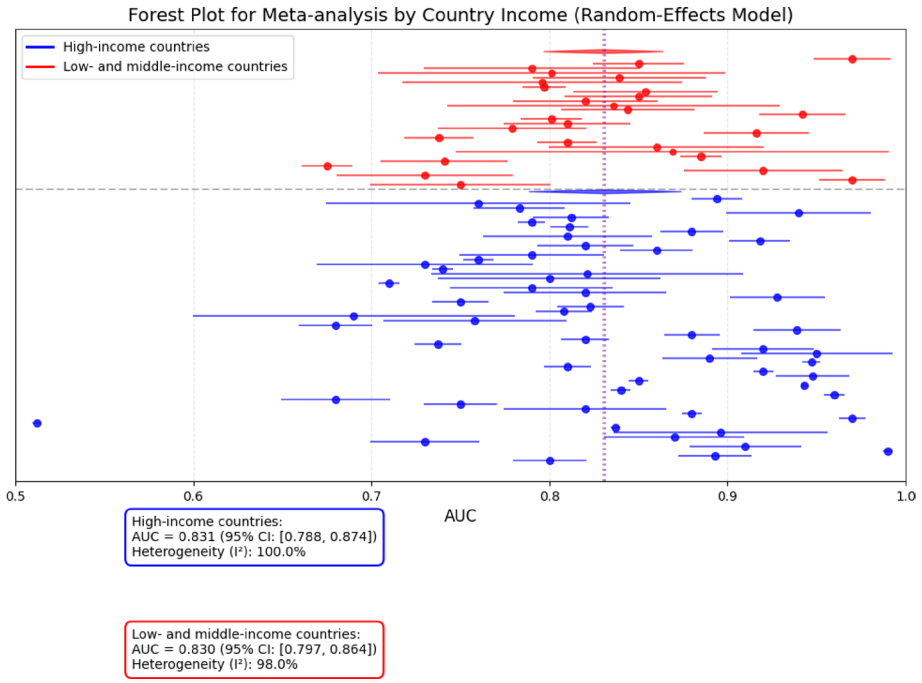

Fig. 9. Forest plot comparing AUC values by country income level. Forest plot comparing pooled AUC values between high-income countries (AUC 0.831, 95% CI 0.788–0.874) and low- and middle-income countries (AUC 0.830, 95% CI 0.797–0.864).

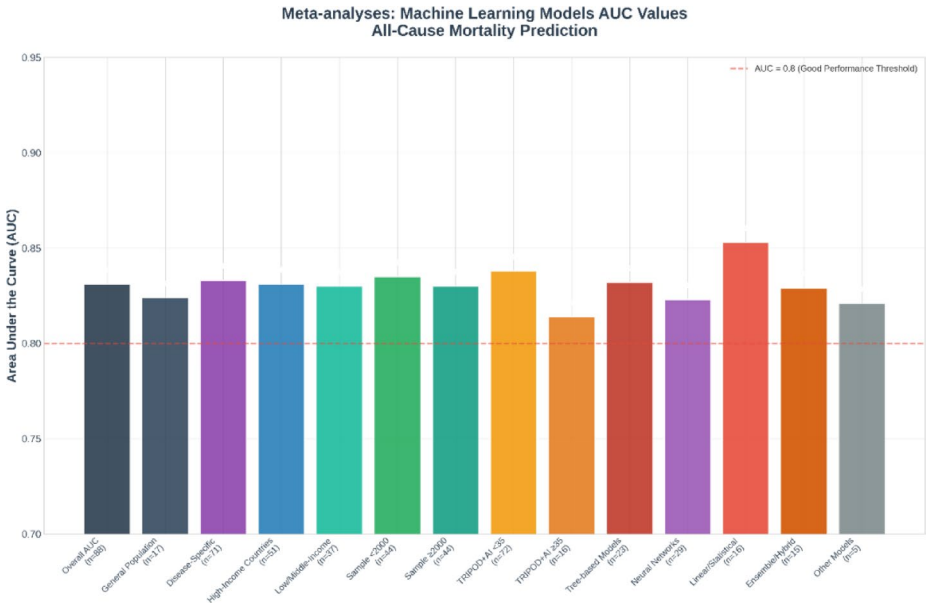

Fig. 10. Summary of pooled AUC values for all-cause mortality prediction. Bar chart showing pooled AUC values with 95% confidence intervals for the overall analysis and subgroup comparisons (general vs. disease-specific populations, and high-income vs. low/middle-income countries).

standardized clinical trajectories and stronger biomarkers (e.g., NT-proBNP and troponin in heart failure patients[30], general populations present multiple mortality pathways). Our findings indicated that well-designed algorithms can effectively handle this heterogeneity, contradicting our initial hypothesis that disease-specific contexts would present superior predictive performance. This suggests machine learning's robustness across diverse population contexts when properly developed.

The comparable performance across different population types represents an important methodological insight for developing universal prediction models. While Shah et al. (2019) noted the challenges of heterogeneity

Study	Title	Abstract	Background			Objectives	Data		Participants			Data preparation	Outcome		
	1	2	3a	3b	3c	4	5a	5b	6a	6b	6c	7	8a	8b	8c
Banerjee et al., 2021	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
Meredith et al., 2002	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Li et al., 2023	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
Barsasella et al., 2022	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Wang Y. et al., 2021	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Díez-Sanmartín et al., 2023	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
Xiong J et al., 2023	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
Lin et al., 2019	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1
Liu et al., 2022	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1
Shi et al., 2012	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
L. Shi, X.C. Wang, Y.S. Wang, 2013	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1
Puddu and Menotti, 2012	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1
Harris et al., 2019	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Takahama et al., 2023	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Arostegui et al., 2018	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Jing et al., 2022	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Tedesco et al., 2021	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Sakr et al., 2017	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Jones et al., 2021	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Singh et al., 2022	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Lu et al., 2019	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Siegersma et al., 2022	1	1	1	0	1	1	1	1	1	1	0	1	1	0	0
Ulloa Cer0 et al., 2021	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Wang et al., 2021	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Mohammad et al., 2022	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0
Valsaraj et al., 2023	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Li et al., 2023	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Zhou et al., 2022	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Giang et al., 2021	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Forte et al., 2021	1	1	1	1	0	1	1	1	1	1	1	0	1	0	0
Bergquist et al., 2021	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
Hernesniemi et al., 2019	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0
Heyman et al., 2021	1	1	1	1	0	1	1	1	1	1	0	0	0	1	1
Qiu et al., 2022	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Niedziela et al., 2021	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0
Mostafaei et al., 2023	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1
Cui et al., 2022	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0
Parikh et al., 2019	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0
Tong et al., 2021	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0
de Capretz et al., 2023	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0

Continued

Study	Title	Abstract	Background			Objectives	Data		Participants		Data preparation	Outcome		
	1	2	3a	3b	3c	4	5a	5b	6a	6b	6c	7	8a	8b	8c
Tian et al., 2023	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0
Mamprin et al., 2021	1	1	1	1	0	1	1	1	1	1	1	0	1	0	1
Motwani et al., 2016	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0
Santos et al., 2019	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0
Feng et al., 2023	1	1	1	1	0	1	1	1	1	1	1	0	1	0	1
Yu et al., 2022	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0
Tamminen et al., 2021	1	1	1	1	0	1	0	1	1	1	0	0	1	0	0
Xu et al., 2023	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Katsiferis et al., 2023	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
Kanda et al., 2022	1	1	1	1	0	1	1	1	1	1	1	0	1	0	0
Lu et al., 2021	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0
Scrutinio et al., 2020	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0
Li et al., 2020	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0
Guo et al., 2022	1	1	1	1	0	1	1	1	1	1	1	0	1	0	0
Wang et al., 2023	1	1	1	1	0	1	1	1	1	1	1	0	1	0	0
Li et al., 2023	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0
Asrian et al., 2024	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0
Ivanics et al., 2023	1	1	1	1	0	1	1	1	1	1		0	1	0	0
Behnoush et al., 2023	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0
Kampaktsis et al., 2023	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Lin et al., 2020	1	1	1	1	0	1	1	0	1	1	0	1	1	0	0
Lin et al., 2023	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Liu et al., 2022	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0
Forssten et al., 2021	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Alimbayev et al., 2023	1	1	1	1	0	1	1	1	1	0	0	1	1	0	0
El-Bouri et al., 2023	1	1	1	1	0	1	1	1	1	0	0	1	1	0	0
Park et al., 2022	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Penso et al., 2021	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Guo et al., 2021	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Abedi et al., 2021	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0

Continued

Study	Title	Abstract	Background			Objectives	Data		Participants			Data preparation	Outcome		
	1	2	3a	3b	3c	4	5a	5b	6a	6b	6c	7	8a	8b	8c
Zhou et al., 2023	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Rauf et al., 2023	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Shi et al., 2022	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Zhou et al., 2021	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Lee et al., 2021	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0
Tran et al., 2023	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
Raghunath et al., 2020	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Kawano et al., 2022	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Weng et al., 2019	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
Zhou et al., 2021	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Huang et al., 2017	1	1	1	1	0	1	1	1	0	1	1	0	1	0	0
Sheng et al., 2020	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Zachariah et al., 2022	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Unterhuber et al., 2021	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Wu et al., 2024	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Hwangbo et al., 2022	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Ross et al., 2016	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0
Wang et al., 2019	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0

Study	Predictors			Sample size	Missing data	Analytical methods							Class imbalance	Fairness	Model output
	9a	9b	9c	10	11	12a	12b	12c	12d	12e	12f.	12 g	13	14	15
Banerjee et al., 2021	1	1	0	1	1	1	1	1	0	1	1	0	1	0	1
Meredith et al., 2002	1	1	0	1	1	1	1	1	1	1	1	0	0	0	1
Li et al., 2023	1	1	0	1	1	1	1	1	1	1	1	0	0	0	1
Barsasella et al., 2022	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1
Wang Y. et al., 2021	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1
Díez-Sanmartín et al., 2023	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1
Xiong J et al., 2023	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1
Lin et al., 2019	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1
Liu et al., 2022	1	1	0	1	1	1	1	1	1	1	0	1	1	0	1
Shi et al., 2012	1	1	0	1	1	1	1	1	0	1	0	1	0	0	1
L. Shi, X.C. Wang, Y.S. Wang, 2013	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1

www.nature.com/scientificreports/

Study	Predictors			Sample size	Missing data	Analytical methods							Class imbalance	Fairness	Model output
	9a	9b	9c	10	11	12a	12b	12c	12d	12e	12f.	12 g	13	14	15
Puddu and Menotti, 2012	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1
Harris et al., 2019	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1
Takahama et al., 2023	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1
Arostegui et al., 2018	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1
Jing et al., 2022	1	1	0	1	0	1	1	1	0	1	0	0	0	0	1
Tedesco et al., 2021	1	1	0	1	1	1	1	1	1	1	0	1	1	0	1
Sakr et al., 2017	1	1	0	1	1	1	1	0	1	1	0	1	1	0	1
Jones et al., 2021	1	1	0	1	1	1	1	1	1	1	0	1	1	0	1
Singh et al., 2022	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1
Lu et al., 2019	0	1	0	1	1	1	1	1	0	1	0	0	0	0	1
Siegersma et al., 2022	0	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Ulloa Cer0 et al., 2021	0	1	0	1	1	1	1	1	0	1	0	0	1	0	1
Wang et al., 2021	0	1	0	0	0	1	1	1	0	1	0	0	1	0	1
Mohammad et al., 2022	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Valsaraj et al., 2023	1	1	0	0	1	1	1	1	0	1	0	0	1	0	1
Li et al., 2023	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Zhou et al., 2022	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Giang et al., 2021	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Forte et al., 2021	1	1	0	1	1	1	0	1	0	1	0	1	0	0	1
Bergquist et al., 2021	0	0	0	1	0	1	0	0	0	1	1	1	0	1	1
Hernesniemi et al., 2019	1	1	0	1	1	1	1	1	0	1	0	1	1	0	1
Heyman et al., 2021	1	1	1	1	1	1	0	1	0	1	0	1	1	0	1
Qiu et al., 2022	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1
Niedziela et al., 2021	1	1	0	1	1	1	1	0	0	1	0	1	1	0	1
Mostafaei et al., 2023	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1
Cui et al., 2022	1	1	0	1	1	1	0	1	0	1	0	1	0	0	1
Parikh et al., 2019	1	1	0	1	1	1	1	1	0	1	0	1	0	0	1
Tong et al., 2021	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0
de Capretz et al., 2023	1	1	0	1	1	1	0	1	0	1	0	1	0	0	1
Tian et al., 2023	1	1	0	1	1	1	0	1	1	1	0	0	0	1	1
Mamprin et al., 2021	1	1	0	1	1	1	1	1	1	1	0	1	1	0	1

Study	Predictors			Sample size	Missing data	Analytical methods							Class imbalance	Fairness	Model output
	9a	9b	9c	10	11	12a	12b	12c	12d	12e	12f.	12 g	13	14	15
Motwani et al., 2016	1	1	1	1	0	1	0	0	0	1	0	1	0	0	1
Santos et al., 2019	1	1	0	1	1	1	1	1	0	1	0	1	0	0	1
Feng et al., 2023	1	1	0	1	0	1	0	0	0	1	0	1	0	0	1
Yu et al., 2022	1	1	0	1	1	1	0	1	0	1	0	1	0	0	1
Tamminen et al., 2021	1	1	0	0	1	1	0	0	0	1	0	1	0	0	1
Xu et al., 2023	1	1	0	1	1	1	1	1	0	1	0	1	0	1	1
Katsiferis et al., 2023	1	1	0	1	1	1	0	1	0	1	0	1	0	1	1
Kanda et al., 2022	1	1	0	1	1	1	0	1	1	1	0	1	0	0	1
Lu et al., 2021	1	1	0	1	0	1	0	1	0	1	0	1	0	0	1
Scrutinio et al., 2020	1	1	1	0	0	1	0	1	0	1	0	1	1	0	1
Li et al., 2020	1	1	0	0	1	0	1	1	0	1	0	1	0	0	1
Guo et al., 2022	1	1	1	1	0	1	0	1	0	1	0	1	0	0	1
Wang et al., 2023	1	1	0	1	1	1	1	1	0	1	0	1	1	0	1
Li et al., 2023	1	1	0	0	0	1	0	1	0	1	0	1	0	0	1
Asrian et al., 2024	1	1	0	1	1	1	0	0	0	1	0	1	0	0	1
Ivanics et al., 2023	1	1	0	1	1	1	1	0	1	1	0	1	0	0	1
Behnoush et al., 2023	1	1	0	1	1	1	0	1	0	1	0	1	1	0	1
Kampaktsis et al., 2023	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Lin et al., 2020	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Lin et al., 2023	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Liu et al., 2022	1	1	0	0	1	1	1	1	0	1	0	0	1	0	1
Forssten et al., 2021	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Alimbayev et al., 2023	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
El-Bouri et al., 2023	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Park et al., 2022	1	1	0	0	1	1	1	1	1	1	0	0	0	0	1
Penso et al., 2021	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Guo et al., 2021	1	1	0	0	1	1	1	1	0	1	0	0	1	0	1
Abedi et al., 2021	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Zhou et al., 2023	1	1	0	0	0	1	1	1	0	1	0	0	0	0	1
Rauf et al., 2023	1	1	0	0	0	1	0	1	0	1	0	0	0	0	1
Shi et al., 2022	1	1	0	0	0	1	0	1	0	1	0	0	0	0	1

Study	Predictors			Sample size	Missing data	Analytical methods								Class imbalance	Fairness	Model output
	9a	9b	9c	10	11	12a	12b	12c	12d	12e	12f.	12 g	13	14	15
Zhou et al., 2021	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
Lee et al., 2021	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Tran et al., 2023	1	1	0	0	1	1	0	1	0	1	0	0	0	0	1
Raghunath et al., 2020	1	1	0	0	1	1	1	1	1	1	0	0	1	0	1
Kawano et al., 2022	1	1	0	0	1	1	1	1	1	1	0	0	1	0	1
Weng et al., 2019	1	1	0	0	1	1	0	1	0	1	0	0	0	0	1
Zhou et al., 2021	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Huang et al., 2017	1	1	0	0	0	1	0	1	0	1	0	0	0	0	1
Sheng et al., 2020	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Zachariah et al., 2022	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Unterhuber et al., 2021	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1
Wu et al., 2024	1	1	0	0	1	1	0	1	0	1	0	0	0	0	1
Hwangbo et al., 2022	1	1	0	0	1	1	0	1	0	1	0	0	0	0	1
Ross et al., 2016	1	1	0	0	1	1	0	1	0	1	0	0	0	0	1
Wang et al., 2019	0	1	1	0	0	1	1	0	1	0	1	0	0	0	1

Study	Training versus evaluation	Ethical approval	Funding	Conflict of interest	Protocol	Registration	Data sharing	Code sharing	Patient and public involvement	Participants			Model development	Model specification
	16	17	18a	18b	18c	18d	18e	18f.	19	20a	20b	20c	21	22
Banerjee et al., 2021	1	1	0	0	0	0	1	1	0	1	1	1	1	0
Meredith et al., 2002	1	1	0	1	0	0	0	0	0	1	1	0	1	1
Li et al., 2023	1	1	1	1	0	0	1	0	0	1	1	1	1	1
Barsasella et al., 2022	1	1	1	1	0	0	1	0	0	1	1	1	1	1
Wang Y. et al., 2021	1	1	1	1	0	0	1	0	0	1	1	1	1	1
Díez-Sanmartín et al., 2023	1	1	1	1	0	0	1	0	0	1	1	1	1	1
Xiong J et al., 2023	1	1	1	1	0	0	1	0	0	1	1	1	1	1
Lin et al., 2019	1	1	1	1	1	0	1	0	0	1	1	0	1	1
Liu et al., 2022	1	1	1	1	1	0	1	0	0	1	1	0	1	1
Shi et al., 2012	1	1	1	0	0	0	0	0	0	1	1	0	1	1
L. Shi, X.C. Wang, Y.S. Wang, 2013	1	1	1	0	0	0	0	0	0	1	1	0	1	1
Puddu and Menotti, 2012	1	1	1	1	0	0	1	0	0	1	1	1	1	1

Study	Training versus evaluation	Ethical approval	Funding	Conflict of interest	Protocol	Registration	Data sharing	Code sharing	Patient and public involvement	Participants			Model development	Model specification	
	16	17	18a	18b	18c	18d	18e	18f.	19	20a	20b	20c	21	22	
Harris et al., 2019	1	1	1	1	0	0	0	0	0	1	1	1	1	1	
Takahama et al., 2023	1	1	1	1	1	1	1	1	0	1	1	1	1	1	
Arostegui et al., 2018	1	1	1	1	0	1	1	0	0	1	1	1	1	1	
Jing et al., 2022	1	0	1	1	0	0	0	0	0	1	1	0	1	0	
Tedesco et al., 2021	1	1	1	1	0	1	0	0	0	1	1	1	1	1	
Sakr et al., 2017	1	1	1	1	0	1	0	0	0	1	1	1	1	1	
Jones et al., 2021	1	1	1	1	0	1	0	0	0	1	1	1	1	1	
Singh et al., 2022	1	1	1	1	0	1	0	0	0	1	1	1	1	1	
Lu et al., 2019	1	1	1	1	1	0	1	1	0	1	1	0	1	0	
Siegersma et al., 2022	1	1	1	1	0	0	1	1	0	1	1	0	1	0	
Ulloa Cer0 et al., 2021	1	1	1	1	0	0	1	1	0	1	1	0	1	0	
Wang et al., 2021	0	1	1	1	0	0	0	1	0	1	1	0	1	0	
Mohammad et al., 2022	1	1	1	1	0	0	1	1	0	1	1	0	1	0	
Valsaraj et al., 2023	1	1	1	1	0	0	1	1	0	1	1	0	1	0	
Li et al., 2023	1	1	1	1	0	0	1	0	0	1	1	0	1	0	
Zhou et al., 2022	0	1	1	1	0	0	1	0	0	1	1	0	1	0	
Giang et al., 2021	0	1	1	1	0	0	1	0	0	1	1	0	1	0	
Forte et al., 2021	0	1	0	1	0	0	0	0	0	1	0	0	1	0	
Bergquist et al., 2021	1	0	1	1	0	0	0	1	1	1	1	1	1	0	
Hernesniemi et al., 2019	0	0	1	1	0	0	0	0	0	1	0	0	1	0	
Heyman et al., 2021	1	0	0	0	0	0	0	0	1	1	0	1	1	0	
Qiu et al., 2022	1	1	1	1	0	0	1	1	0	1	1	1	1	0	
Niedziela et al., 2021	0	1	0	1	0	0	0	0	0	1	0	0	0	0	
Mostafaei et al., 2023	1	1	1	1	0	0	1	0	0	1	0	1	1	0	
Cui et al., 2022	0	1	1	1	0	0	1	1	0	1	1	1	1	0	
Parikh et al., 2019	1	1	1	1	0	0	1	1	0	1	1	0	0	0	
Tong et al., 2021	1	1	1	1	0	0	1	0	0	1	0	1	1	0	
de Capretz et al., 2023	0	1	1	1	0	0	1	0	0	1	1	1	1	0	
Tian et al., 2023	0	1	1	1	0	0	1	0	0	1	1	0	0	0	
Mamprin et al., 2021	1	1	1	1	0	0	1	0	0	1	1	1	1	0	

Study	Training versus evaluation	Ethical approval	Funding	Conflict of interest	Protocol	Registration	Data sharing	Code sharing	Patient and public involvement	Participants			Model development	Model specification	
	16	17	18a	18b	18c	18d	18e	18f.	19	20a	20b	20c	21	22	
Motwani et al., 2016	0	1	1	1	0	0	0	0	0	1	0	0	0	0	
Santos et al., 2019	1	1	1	0	0	0	0	0	0	0	0	0	1	0	
Feng et al., 2023	0	1	1	1	0	0	0	0	0	1	0	0	0	0	
Yu et al., 2022	0	1	1	1	0	0	1	0	0	1	1	0	0	0	
Tamminen et al., 2021	0	1	1	1	0	0	0	0	0	1	0	0	0	0	
Xu et al., 2023	1	1	1	1	0	0	1	1	0	1	1	1	1	0	
Katsiferis et al., 2023	1	1	1	1	0	0	1	1	0	1	1	1	1	0	
Kanda et al., 2022	1	1	1	1	0	0	1	0	0	1	1	1	1	0	
Lu et al., 2021	0	1	0	1	0	0	0	0	0	1	1	0	0	0	
Scrutinio et al., 2020	0	1	0	1	0	0	1	0	0	1	0	0	0	0	
Li et al., 2020	0	1	1	1	0	0	0	0	0	1	1	0	0	0	
Guo et al., 2022	0	1	1	1	0	0	1	0	0	1	1	0	0	1	
Wang et al., 2023	0	1	1	1	0	0	1	0	0	1	1	1	1	0	
Li et al., 2023	0	1	1	1	0	0	1	0	0	1	1	1	1	1	
Asrian et al., 2024	0	0	1	0	0	0	0	0	0	1	0	0	0	0	
Ivanics et al., 2023	1	1	0	1	0	0	1	0	0	1	0	1	1	0	
Behnoush et al., 2023	0	1	0	1	0	0	1	1	0	1	0	1	0	0	
Kampaktsis et al., 2023	0	1	1	1	0	0	1	0	0	1	1	0	1	0	
Lin et al., 2020	1	1	1	1	0	0	0	0	0	1	1	0	1	0	
Lin et al., 2023	0	1	1	1	0	0	1	0	0	1	1	0	1	0	
Liu et al., 2022	1	1	1	1	0	0	1	1	0	1	1	0	1	0	
Forssten et al., 2021	0	1	1	1	0	0	1	0	0	1	1	0	1	0	
Alimbayev et al., 2023	0	1	1	1	0	0	1	0	0	1	1	0	1	0	
El-Bouri et al., 2023	0	1	1	1	0	0	1	0	0	1	1	0	1	0	
Park et al., 2022	1	1	0	0	0	1	1	0	0	1	1	0	1	0	
Penso et al., 2021	0	1	1	1	0	0	1	1	0	1	1	0	1	0	
Guo et al., 2021	0	1	1	1	0	0	1	0	0	1	1	0	1	0	
Abedi et al., 2021	0	1	1	1	0	0	1	1	0	1	1	0	1	0	
Zhou et al., 2023	0	1	0	1	0	0	1	0	0	1	1	0	1	0	
Rauf et al., 2023	0	1	0	1	0	0	1	0	0	1	1	0	1	0	

Study	Training versus evaluation	Ethical approval	Funding	Conflict of interest	Protocol	Registration	Data sharing	Code sharing	Patient and public involvement	Participants			Model development	Model specification	
	16	17	18a	18b	18c	18d	18e	18f.	19	20a	20b	20c	21	22	
Shi et al., 2022	0	1	1	1	0	0	1	0	0	1	1	0	1	0	
Zhou et al., 2021	1	1	1	1	0	0	1	0	0	1	1	0	1	0	
Lee et al., 2021	0	1	1	1	0	0	0	0	0	1	1	0	1	0	
Tran et al., 2023	0	1	1	1	0	0	1	0	0	1	1	0	1	0	
Raghunath et al., 2020	0	1	1	1	0	0	1	1	0	1	1	1	1	0	
Kawano et al., 2022	0	1	1	1	0	0	1	1	0	1	1	1	1	0	
Weng et al., 2019	0	1	1	1	0	0	1	1	0	1	1	0	1	0	
Zhou et al., 2021	1	1	1	0	0	0	0	0	0	1	1	0	1	0	
Huang et al., 2017	1	1	1	1	0	0	1	0	0	1	1	0	1	0	
Sheng et al., 2020	1	1	1	0	0	0	0	0	0	1	1	0	1	0	
Zachariah et al., 2022	1	1	1	0	0	0	0	0	0	1	1	0	1	0	
Unterhuber et al., 2021	1	1	1	1	0	0	1	1	0	1	1	0	1	1	
Wu et al., 2024	0	1	1	1	0	1	1	0	0	1	1	0	1	0	
Hwangbo et al., 2022	0	1	1	1	0	1	1	0	0	1	1	0	1	0	
Ross et al., 2016	0	1	1	1	0	0	0	0	0	1	1	0	1	0	
Wang et al., 2019	0	1	1	1	0	0	0	0	0	1	1	0	1	0	

Study	Model performance		Model updating	Interpretation	Limitations	Usability of the model in the context of current care			Final score	
	23a	23b	24	25	26	27a	27b	27c		
Banerjee et al., 2021	1	0	0	1	1	0	0	1	35	
Meredith et al., 2002	1	0	0	1	1	0	0	1	33	
Li et al., 2023	1	1	0	1	1	0	0	1	38	
Barsasella et al., 2022	1	1	0	1	1	0	0	1	38	
Wang Y. et al., 2021	1	1	0	1	1	0	0	1	38	
Díez-Sanmartín et al., 2023	1	0	1	1	1	0	0	1	39	
Xiong J et al., 2023	1	1	0	1	1	0	0	1	39	
Lin et al., 2019	1	1	0	1	1	0	0	1	39	
Liu et al., 2022	1	1	0	1	1	1	0	1	41	
Shi et al., 2012	1	0	0	1	1	0	0	1	33	

Study	Model performance		Model updating	Interpretation	Limitations	Usability of the model in the context of current care			Final score	
	23a	23b	24	25	26	27a	27b	27c		
L. Shi, X.C. Wang, Y.S. Wang, 2013	1	1	0	1	1	0	0	1	36	
Puddu and Menotti, 2012	1	1	0	1	1	0	0	1	40	
Harris et al., 2019	1	1	0	1	1	1	0	1	40	
Takahama et al., 2023	1	1	1	1	1	1	0	1	45	
Arostegui et al., 2018	1	1	0	1	1	1	0	1	42	
Jing et al., 2022	1	1	0	1	1	0	0	1	30	
Tedesco et al., 2021	1	1	1	1	1	1	0	1	41	
Sakr et al., 2017	1	1	1	1	1	1	0	1	40	
Jones et al., 2021	1	1	1	1	1	1	0	1	41	
Singh et al., 2022	1	1	1	1	1	1	0	1	42	
Lu et al., 2019	1	0	0	1	1	0	0	1	33	
Siegersma et al., 2022	1	0	0	1	1	0	0	1	31	
Ulloa Cer0 et al., 2021	1	0	0	1	1	0	0	1	33	
Wang et al., 2021	0	0	1	1	0	0	1	1	29	
Mohammad et al., 2022	1	0	0	1	1	0	0	1	33	
Valsaraj et al., 2023	1	0	0	1	1	0	0	1	33	
Li et al., 2023	1	0	0	1	1	0	0	1	31	
Zhou et al., 2022	1	0	0	1	1	0	0	1	31	
Giang et al., 2021	1	0	1	1	0	0	0	1	31	
Forte et al., 2021	1	0	0	0	1	0	0	1	27	
Bergquist et al., 2021	1	0	0	1	1	0	0	1	32	
Hernesniemi et al., 2019	0	0	0	0	1	0	0	1	27	
Heyman et al., 2021	0	0	0	1	1	1	0	1	31	
Qiu et al., 2022	1	0	0	0	1	0	1	1	36	
Niedziela et al., 2021	0	0	0	0	1	0	0	1	25	
Mostafaei et al., 2023	0	1	1	1	1	0	0	1	38	
Cui et al., 2022	0	0	0	0	1	0	0	1	30	
Parikh et al., 2019	0	0	0	0	1	0	0	1	30	
Tong et al., 2021	0	1	0	0	1	0	0	1	25	

Study	Model performance		Model updating	Interpretation	Limitations	Usability of the model in the context of current care			Final score	
	23a	23b	24	25	26	27a	27b	27c		
de Capretz et al., 2023	0	0	0	0	1	0	0	1	30	
Tian et al., 2023	1	1	0	1	1	0	0	1	31	
Mamprin et al., 2021	0	1	0	0	1	1	0	1	37	
Motwani et al., 2016	0	0	0	0	1	0	0	1	24	
Santos et al., 2019	0	0	0	0	1	0	0	1	27	
Feng et al., 2023	1	0	0	0	1	0	0	1	26	
Yu et al., 2022	0	0	0	1	1	0	0	1	28	
Tamminen et al., 2021	1	0	0	0	1	0	0	1	23	
Xu et al., 2023	1	0	0	1	1	0	0	1	36	
Katsiferis et al., 2023	1	0	0	1	1	0	0	1	36	
Kanda et al., 2022	0	1	0	0	1	0	0	1	33	
Lu et al., 2021	0	0	0	0	1	0	0	1	25	
Scrutinio et al., 2020	0	0	0	0	1	0	0	1	25	
Li et al., 2020	0	0	0	0	1	0	0	0	24	
Guo et al., 2022	0	0	0	0	1	0	0	1	29	
Wang et al., 2023	0	0	0	0	1	0	0	1	32	
Li et al., 2023	0	0	0	0	1	0	1	1	30	
Asrian et al., 2024	0	0	1	0	1	0	0	1	24	
Ivanics et al., 2023	1	1	0	1	1	1	0	1	33	
Behnoush et al., 2023	1	0	0	1	1	0	0	1	31	
Kampaktsis et al., 2023	1	0	0	1	1	0	0	1	31	
Lin et al., 2020	1	0	0	1	1	0	0	1	29	
Lin et al., 2023	1	0	0	1	1	0	0	1	31	
Liu et al., 2022	1	0	0	1	1	0	0	1	34	
Forssten et al., 2021	1	0	0	1	1	0	0	1	30	
Alimbayev et al., 2023	1	0	0	1	1	0	0	1	29	
El-Bouri et al., 2023	1	0	0	1	1	0	0	1	29	
Park et al., 2022	1	0	0	1	1	0	0	1	31	
Penso et al., 2021	1	0	0	1	1	0	0	1	31	

Study	Model performance		Model updating	Interpretation	Limitations	Usability of the model in the context of current care			Final score						
	23a	23b	24	25	26	27a	27b	27c							
Guo et al., 2021	1	0	0	1	1	0	0	1	31						
Abedi et al., 2021	1	0	0	1	1	0	0	1	31						
Zhou et al., 2023	1	0	1	1	0	0	0	1	28						
Rauf et al., 2023	1	0	0	1	1	0	0	1	28						
Shi et al., 2022	1	0	0	1	1	0	0	1	29						
Zhou et al., 2021	1	0	0	1	1	0	0	1	38						
Lee et al., 2021	1	0	1	1	1	0	0	1	32						
Tran et al., 2023	1	0	1	1	0	0	0	1	30						
Raghunath et al., 2020	1	1	0	1	1	0	0	1	35						
Kawano et al., 2022	1	1	0	1	1	0	0	1	35						
Weng et al., 2019	1	1	1	1	1	0	0	1	34						
Zhou et al., 2021	1	0	0	1	1	0	0	1	29						
Huang et al., 2017	1	0	0	1	1	0	0	1	28						
Sheng et al., 2020	1	0	1	1	0	0	0	1	29						
Zachariah et al., 2022	1	0	0	1	1	0	0	1	29						
Unterhuber et al., 2021	1	1	0	1	1	0	0	1	34						
Wu et al., 2024	1	0	0	1	1	0	0	1	30						
Hwangbo et al., 2022	1	0	0	1	1	0	0	1	30						
Ross et al., 2016	1	0	0	1	1	0	0	1	28						
Wang et al., 2019	1	0	0	1	1	0	0	1	29						

Table 2. Item-by-item quality assessment using TRIPOD + AI criteria (n = 88). Comprehensive quality assessment results showing compliance with each of the 27 TRIPOD + AI reporting criteria across all included studies, with final quality scores ranging from 23 to 45 points.

Variable	AUC (95%CI)	I^2 (%)
Sample size		
Less than 2000	0.835 (0.799–0.871)	96.1
2000 or more	0.830 (0.789–0.870)	100
TRIPOD + AI		
Less than 35 points	0.838 (0.817–0.858)	99.8
35 or more	0.814 (0.704–0.924)	100
Models		
Tree-based models	0.832 (0.774–0.890)	100
Neural networks	0.823 (0.793–0.854)	99.7
Linear/statistical	0.853 (0.759–0.947)	99.2
Ensemble/hybrid models	0.829 (0.781–0.878)	98
Other	0.821 (0.734–0.907)	98.3
Imputation of confidence intervals		
Non-imputed	0.815 (0.768–0.862)	100
I mputed	0.856 (0.823–0.890)	99.4

Table 3. Subgroup analysis results by study characteristics and model types. Pooled AUC values with 95% confidence intervals and heterogeneity measures (I^2) for subgroup analyses by sample size, TRIPOD + AI quality score, machine learning model categories, and confidence interval imputation status. *Note* Other models category includes algorithms used by single studies: DSM (n = 1), DLS-MSM (n = 1), ICISS (n = 1), Support Vector Machine (n = 1), and Bayesian Network (n = 1).

Variable	Comparison	β coefficient	Standard error	*p*-value
Country income level	High-income versus low/middle-income	0.0561	0.1041	0.591
Population type	General versus disease-specific	− 0.1919	0.0257	< 0.001
Sample size	≥ 2000 vs < 2000 participants	0.0402	0.1609	0.803
TRIPOD + AI score	≥ 35 vs < 35 points	− 0.1341	0.0312	< 0.001
Algorithm type	Neural Networks versus tree-based	0.1355	0.0267	< 0.001
Algorithm type	Linear/statistical versus tree-based	− 0.1067	0.1097	0.3334
Algorithm type	Ensemble/hybrid versus tree-based	− 0.0043	0.1057	0.9674
Algorithm type	Other models versus tree-based	− 0.0740	0.2242	0.7421
Confidence interval	Imputed versus non-imputed CI	0.1090	0.0410	0.0093
Outcome prevalence	20–39% versus 0–19%	0.0818	0.0425	0.0578
Outcome prevalence	40% or more versus 0–19%	− 0.0253	0.0735	0.7318

Table 4. Meta-regression analysis of factors associated with AUC performance. Results of univariate meta-regression analyses examining the relationship between study characteristics (country income level, population type, sample size, quality score, algorithm type, confidence interval imputation, and outcome prevalence) and AUC performance, showing β coefficients, standard errors, and *p*-values. *Note*β coefficient represents the difference in AUC between comparison groups; Reference categories: Low/Middle-income countries, Disease-specific populations, < 2000 participants, < 35 TRIPOD + AI points, Tree-based models; Positive β indicates higher AUC in the first group; negative β indicates lower AUC in the first group; Other models category includes algorithms used by single studies: DSM (n = 1), DLS-MSM (n = 1), ICISS (n = 1), Support Vector Machine (n = 1), and Bayesian Network (n = 1).

in clinical prediction modeling[31], our results suggest that these challenges can be overcome with robust modeling strategies such as ensemble methods that can handle diverse risk patterns, comprehensive feature engineering that captures population-specific risk factors, and stratified validation approaches that ensure consistent performance across different demographic groups.

The lack of consistency in predictor variables significantly affects the generalizability and equity of machine learning models for all-cause mortality prediction. Demographic (97.73%) and clinical (88.64%) variables were most commonly used, reflecting their accessibility and strong mortality associations. Laboratory (47.73%) and imaging (20.45%) variables were less frequent, with imaging predominantly used in high-income countries due to advanced infrastructure (e.g., Siegersma et al., 2022, achieved AUC 0.96 with ECGs[26]). Socioeconomic and behavioral factors were underutilized (27.27% of studies) despite their relevance to health equity. This predictor variability contributes to the observed heterogeneity and raises concerns about model applicability in different settings. Future studies should adopt standardized minimum reporting requirements, including sociodemographic variables and standardized morbidity indices. Establishing consensus-based predictor sets

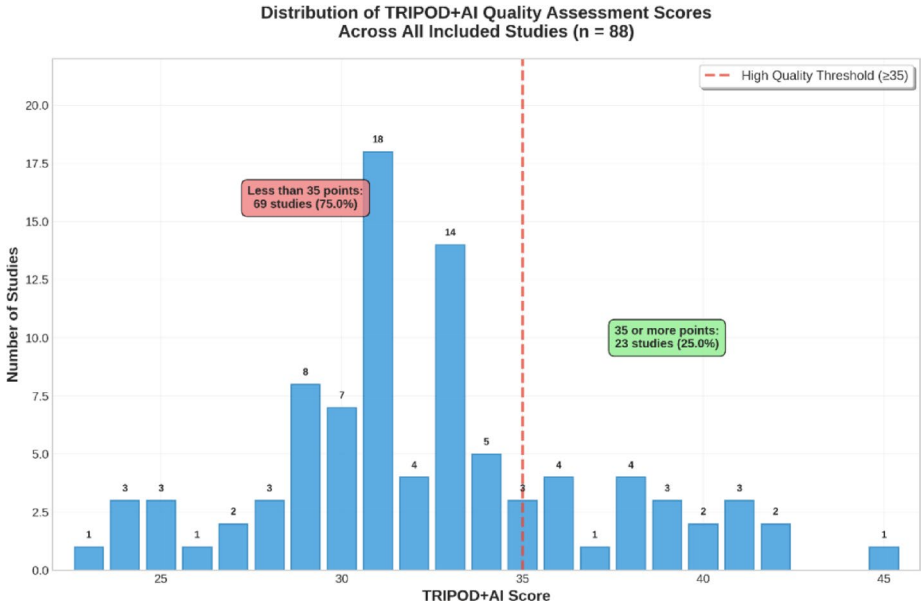

Fig. 11. Distribution of TRIPOD + AI quality assessment scores. Histogram showing the distribution of TRIPOD + AI quality assessment scores across all 88 included studies, with scores ranging from 23 to 45 points and most studies clustering around 31–33 points.

Category	n (%) of studies
Inclusion of social variables*	
Race/Ethnicity	3 (3.4%)
Education	3 (3.4%)
Income/Socioeconomic Status	9 (10.2%)
No social variables	79 (89.8%)
Demographic diversity reported	
Complete racial composition	3 (3.4%)
Age and sex only	76 (86.4%)
Limited demographic data	9 (10.2%)
Subgroup analysis by	
Race/ethnicity	0 (0.0%)
Sex/gender	2 (2.3%)
Age	5 (5.7%)
Socioeconomic status	0 (0.0%)
External validation	
Multiple/international populations	7 (8.0%)
Internal validation only	81 (92.0%)

Table 5. Assessment of equity-related reporting across included studies. Quantitative summary of equity considerations including the inclusion of social determinant variables, demographic diversity reporting, subgroup analyses by demographic characteristics, and external validation practices across the 88 included studies. *The sum is more than 100% because the same study can appear more than once per category.

for different population types (general vs. disease-specific) would improve comparability and facilitate model validation across settings.

Country-income-stratified analysis revealed similar performance of mortality prediction models across economic contexts. Models from high-income countries achieved virtually equal AUC values to those from low- and middle-income countries, both higher than 0.8. While high-income countries more commonly conduct large-scale studies with extended follow-up (e.g., UK Biobank[32]), our results show that studies from low- and middle-income countries achieve comparable performance. The specific factors contributing to this similarity require further investigation due to the found heterogeneity among studies[32].

To maintain and further strengthen the observed equity in model performance across different economic contexts, we propose additional strategies and techniques associated with equity-focused analysis. First,

establishing data-sharing frameworks between institutions in high-, middle-, and low-income countries could improve the representativeness and size of datasets across all contexts. Recent initiatives have demonstrated the feasibility of standardized global data collection[32], but there is room to improve equitable global participation and representativeness. Second, model transfer techniques could adapt models developed in one context to another by adjusting for local epidemiological characteristics. Wiens et al.[33] demonstrated that transfer learning approaches can successfully adapt clinical prediction models across hospitals with varying resources and patient populations. Additionally, federated machine learning can enable collaborative model development without direct sharing of sensitive data between institutions[34]. Chen et al.[35] proposed a framework for evaluating algorithmic fairness in clinical prediction models that could be adapted for all-cause mortality prediction. Such models should be evaluated not only for overall performance but also for equitable performance across population subgroups[36].

The substantial heterogeneity observed across all analyses represents our most significant finding, suggesting considerable variation in model development and study methodology despite similar overall performance metrics. This heterogeneity can be due to differences in predictor variables, modeling approaches, local health systems contexts, and mortality follow-up periods. While subgroup analysis showed overlapping confidence intervals between disease-specific (AUC 0.833) and general populations (AUC 0.824), meta-regression showed a statistically significant advantage for disease-specific populations, suggesting that although the mean difference is modest and descriptively similar, it is statistically consistent across studies.

The high degree of heterogeneity suggests that the model's performance is heavily influenced by the context. Health professionals and stakeholders cannot rely on models to remain accurate across various populations or healthcare environments. While high heterogeneity is commonly observed in meta-analyses including large numbers of studies with diverse methodologies and populations[37–40], this variation suggests that local factors significantly influence model performance. The finding that lower-quality studies reported higher performance suggests potential overfitting and publication bias. For clinical and public health implementation, local validation is required regardless of reported performance.

Studies with lower TRIPOD + AI scores achieved higher AUC values, a counterintuitive finding that can reflect publication bias and overfitting in studies with less rigorous validation strategies. This pattern suggests that studies prioritizing high-performance metrics may result from lower methodological rigor. With only 8% of studies employing external validation, higher methodological quality studies may report more conservative but realistic performance estimates through proper validation process. The interpretation of model performance from poorly reported studies should be carefully done due to overestimation of the true clinical performance as result of potential inadequate validation and potential overfitting. This finding highlights that rigorous study design and validation strategies should be prioritized from inception rather than evaluated post hoc, and suggests the need for mandatory external validation and transparent reporting in future ML mortality prediction studies. The quality-performance paradox observed emphasizes that stringent external validation is essential before any implementation.

Neural networks presented higher performance than tree-based models according to meta-regression analysis. Deep learning models may better capture complex nonlinear relationships to predict mortality. The superior performance of neural networks can reflect their ability to deal with complex interactions between multiple mortality risk factors simultaneously, including unexpected data patterns that traditional algorithms might miss. However, higher performance must be balanced with interpretability concerns in clinical settings. Considering the slight (besides statistically significant) difference, simpler models are the best choice in clinical and health system settings. For implementation, its recommended to prioritize interpretable models (tree-based, logistic regression) when the performance difference is modest (< 0.03 AUC)[41], especially in resource-constrained settings where model transparency is relevant for health professionals' acceptance and regulatory approval. Practical implementation in resource-limited settings faces challenges. However, our findings show comparable performance between LMIC and high-income countries (AUC 0.830 vs. 0.831), suggesting feasible implementation. Resource-constrained settings should prioritize simpler, interpretable models requiring minimal computational power and enabling local validation by clinical staff.

In general, our results indicate a good performance among different algorithms. The quality of data, pre-processing techniques, and representativeness of the target population seem to be more important than the algorithm itself[42,43]. In this scenario, for example, tree-based models, while showing slightly lower discriminative performance in our analysis, offer greater interpretability through their decision process and feature importance. Thus, the higher algorithmic transparency, especially about how models are making decisions, improves the relevance of simpler models (such as tree-based) to be used to predict mortality and effectiveness evaluation. Healthcare regulatory frameworks should emphasize model interpretability requirements for mortality prediction systems, particularly given that tree-based models offer substantial clinical advantages through transparent decision pathways while maintaining comparable performance to more complex approaches, as demonstrated by the modest performance differences observed in our analysis and supported by interpretability research in high-stakes medical applications[44].

An important consideration in the implementation of machine learning models is the trade-off between interpretability and predictive performance. Although neural networks achieved statistically higher performance relative to tree-based models, the observed difference was marginal, with a mean gain of less than 0.02 AUC points. Tree-based models offer substantial advantages in clinical settings through their inherent interpretability, allowing clinicians to understand decision pathways and feature importance rankings[44,45]. This transparency is relevant for the model's adoption, particularly in decisions where model errors can have severe consequences[44]. Given the comparable performance across model types and the critical importance of interpretability in public health applications[46], simpler, more transparent models may be preferable for routine implementation despite

virtually lower discriminative performance. Regulatory approval and clinical adoption can favor transparent models that enable audit trails and staff understanding.

Although populations and economic settings differ substantially, our analysis indicates that machine learning models for mortality prediction reach a comparable level of discriminative accuracy, but often rely on distinct mechanisms to do so. This observation of "multiple routes to success" highlights the need for closer investigation. Future research could explore which specific pre-processing development factors, algorithms, or predictor variable selection contribute to model performance in different contexts.

The substantial number of publications examined and the adherence to methodological guidelines like PROSPERO, PRISMA, and TRIPOD + AI are two of this study's strong points. There are certain restrictions, though. The significant heterogeneity found suggests a great deal of variation in the populations and methods examined, which may have an impact on how broadly applicable our findings are.

An important methodological limitation of our review is the insufficient external validation among the included studies. Only a minority of the 88 analyzed studies employed independent external datasets to evaluate the generalizability of their models. Most studies relied on internal validation (such as cross-validation or train-test splits within the same population), which may overestimate model performance and limit clinical applicability. This absence of robust external validation represents a significant gap in the field, as models that demonstrate high performance in internal validation frequently show a substantial shift when applied to different populations or clinical contexts[47].

Only 19% of studies included general population samples which can be considered a generalizability limitation. Disease-specific models cannot be extrapolated to populations with diverse risk profiles and heterogeneous clinical trajectories. Models developed in patients with specific diseases can rely on well-defined clinical pathways and established biomarkers. In contrast, general population prediction must identify risk among predominantly healthy individuals presenting different and interconnected risk factors. Disease-specific datasets combined with limited external validation (8% of studies) restrict transferability to population-level applications. Disease-specific models use homogeneous cohorts with standardized markers, while population prediction requires identifying risk in predominantly healthy, diverse populations. This sampling bias limits applicability to public health surveillance and national screening programs.

Precision estimates may have been biased as a result of our use of literature-based estimation techniques to deal with missing confidence intervals. Heterogeneity assessments and pooled estimates may be impacted by this. By inflating apparent consistency or exaggerating heterogeneity, attributed confidence intervals can produce artificially narrow or wide precision estimates, which compromises the validity of our meta-analytic findings.

Our analysis was limited by insufficient data on social and behavioral variables, including education, race/ethnicity, and housing conditions. Future studies should prioritize including these social determinants to improve model accuracy and equity. Equity-focused reporting should be mandatory, with fairness assessments required before clinical implementation. We recommend alignment with established AI ethics guidelines to prevent perpetuating healthcare disparities through algorithmic bias. Funding agencies, academic journals, and regulatory bodies should integrate these equity-focused reporting requirements into research funding criteria, manuscript submission guidelines, and clinical approval processes to ensure systemic implementation across the field. Researchers should conduct stratified performance analysis across demographic subgroups, healthcare systems must require local validation before model deployment, and all stakeholders should establish monitoring systems to track equity outcomes post-implementation.

Thus, our analysis also highlights the potential for biases in training data to perpetuate historical and systemic inequities through the use of these models in health settings, particularly when marginalized populations are underrepresented in training datasets. The similar performance across different economic contexts does not necessarily represent equitable models for all demographic groups within these contexts. Most of the analyzed studies did not report model performance stratified by race/ethnicity, gender, or socioeconomic status, leaving potential disparities undetected. Algorithmic fairness requires intentional design choices and comprehensive evaluation across diverse subpopulations[36,48], which should be prioritized in future studies before clinical implementation and deployment.

In conclusion, this meta-analysis showed the potential of machine learning models to predict all-cause mortality across diverse populations and economic contexts. However, findings derive predominantly from disease-specific populations (81% of studies), with limited evidence for general population applicability. The substantial heterogeneity across all analyses indicates a high degree of variation among studies. Future research should prioritize the development of models specifically for general populations, standardized reporting, inclusive data collection that incorporates social determinants of health such as race/ethnicity and socioeconomic status, and rigorous external validation across diverse populations. Finally, the low percentage of external validation requires caution in the generalizability of results and implementation.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 2 August 2024; Accepted: 30 October 2025
Published online: 27 November 2025

References

1. Avinash, B. S., Srisupattarawanit, T. & Ostermeyer, H. Numerical methods for information tracking of noisy and non-smooth data in large-scale statistics. *J. Eng. Res. Rep.* https://doi.org/10.9734/jerr/2019/v6i416957 (2019).

2. Zhang, J. et al. Guest editorial learning from noisy multimedia data. *IEEE Trans. Multimed* https://doi.org/10.1109/TMM.2022.3159014 (2022).
3. Arain, Z., Iliodromiti, S., Slabaugh, G., David, A. L. & Chowdhury, T. T. Machine learning and disease prediction in obstetrics. *Curr. Res. Physiol.* https://doi.org/10.1016/j.crphys.2023.100099 (2023).
4. Veena, S., Sumanth Reddy, D., Lakshmi Kara, C. & Uday Kiran, K. A. Clinical outcome future prediction with decision tree & naive bayes models, in *Advances in Science and Technology*, Vol. 124. AST (2023).
5. Li, Y., Fan, X., Wei, L., Yang, K. & Jiao, M. The impact of high-risk lifestyle factors on all-cause mortality in the US non-communicable disease population. *BMC Public Health* **23**, 422 (2023).
6. Taneri, P. E. et al. Association between ultra-processed food intake and all-cause mortality: A systematic review and meta-analysis. *Am. J. Epidemiol.* **191**, 1323–1335 (2022).
7. Feng, X., Sarma, H., Seubsman, S. A., Sleigh, A. & Kelly, M. The impact of multimorbidity on all-cause mortality: A longitudinal study of 87,151 thai adults. *Int. J. Public Health* **68**, 1606137 (2023).
8. Takahama, H. et al. Clinical application of artificial intelligence algorithm for prediction of one-year mortality in heart failure patients. *Heart Vessels* **38**, 785–792 (2023).
9. Arostegui, I. et al. Combining statistical techniques to predict postsurgical risk of 1-year mortality for patients with colon cancer. *Clin. Epidemiol.* **10**, 235–251 (2018).
10. Xiong, J. et al. A novel machine learning-based programmed cell death-related clinical diagnostic and prognostic model associated with immune infiltration in endometrial cancer. *Front. Oncol.* **13**, 1224071 (2023).
11. Wang, G. et al. Machine learning-based models for predicting mortality and acute kidney injury in critical pulmonary embolism. *BMC Cardiovasc. Disord.* **23**, 385 (2023).
12. Bacevicius, M. & Paulauskaite-Taraseviciene, A. Machine learning algorithms for raw and unbalanced intrusion detection data in a multi-class classification problem. *Appl. Sci. (Switzerland)* **13**, 7328 (2023).
13. Delpino, F. M. et al. Machine learning for predicting chronic diseases: A systematic review. *Public Health* **205**, 14–25 (2022).
14. Delpino, F. M. et al. Does machine learning have a high performance to predict obesity among adults and older adults? A systematic review and meta-analysis. *Nutr. Metab. Cardiovasc. Dis.* **34**(9), 2034–2045 (2024).
15. Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D. & Tzovara, A. Addressing bias in big data and AI for health care: A call for open science. *Patterns* https://doi.org/10.1016/j.patter.2021.100347 (2021).
16. Matthew, P. *et al.* PRISMA 2020 statement: updated guidelines for reporting systematic reviews and meta analyses. *26th Cochrane Colloquium Santiago Chile* (2019).
17. Collins, G. S. et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
18. Meurer, W. J. & Tolles, J. Logistic regression diagnostics understanding how well a model predicts outcomes. *JAMA J. Am. Med. Assoc.* https://doi.org/10.1001/jama.2016.20441 (2017).
19. 7.7.7.2 Standard errors from confidence intervals and P values: difference measures. https://handbook-5-1.cochrane.org/chapter_7/7_7_7_2_obtaining_standard_errors_from_confidence_intervals_and.htm.
20. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
21. Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. Synth. Methods* **1**, 97–111 (2010).
22. Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. Introduction to meta-analysis. *Introd. Meta-Anal.* https://doi.org/10.1002/9780470743386 (2009).
23. Heyman, E. T. et al. Improving machine learning 30-day mortality prediction by discounting surprising deaths. *J. Emerg. Med.* **61**, 763–773 (2021).
24. Bergquist, T. et al. Evaluation of crowdsourced mortality prediction models as a framework for assessing artificial intelligence in medicine. *J. Am. Med. Inform. Assoc.* **31**, 35–44 (2024).
25. Díez-Sanmartín, C., Cabezuelo, A. S. & Belmonte, A. A. A new approach to predicting mortality in dialysis patients using sociodemographic features based on artificial intelligence. *Artif. Intell. Med.* **136**, 102478 (2023).
26. Siegersma, K. R. et al. Deep neural networks reveal novel sex-specific electrocardiographic features relevant for mortality risk. *Eur. Heart J. Digit. Health* **3**, 245–254 (2022).
27. Barsasella, D. et al. A machine learning model to predict length of stay and mortality among diabetes and hypertension inpatients. *Medicina (Kaunas)* **58**, 1568 (2022).
28. Li, D., Fu, J., Zhao, J., Qin, J. & Zhang, L. A deep learning system for heart failure mortality prediction. *PLoS ONE* **18**, e0276835 (2023).
29. Tedesco, S. et al. Comparison of machine learning techniques for mortality prediction in a prospective cohort of older adults. *Int. J. Environ. Res. Public Health* **18**, 12806 (2021).
30. Tang, W. H. W. et al. Prognostic value of baseline and changes in circulating soluble ST2 levels and the effects of nesiritide in acute decompensated heart failure. *JACC Heart Fail.* **4**, 68–77 (2016).
31. Shah, N. D., Steyerberg, E. W. & Kent, D. M. Big data and predictive analytics: Recalibrating expectations. *JAMA J. Am. Med. Assoc.* https://doi.org/10.1001/jama.2018.56024 (2018).
32. Zerillo, J. A. et al. An international collaborative standardizing a comprehensive patient-centered outcomes measurement set for colorectal cancer. *JAMA Oncol* https://doi.org/10.1001/jamaoncol.2017.0417 (2017).
33. Wiens, J., Guttag, J. & Horvitz, E. Patient risk stratification with time-varying parameters: A multitask learning approach. *J. Mach. Learn. Res.* **17**, 1–23 (2016).
34. Rieke, N. et al. The future of digital health with federated learning. *NPJ Digit. Med.* **3**, 119 (2020).
35. Chen, I. Y., Szolovits, P. & Ghassemi, M. Can AI help reduce disparities in general medical and mental health care?. *AMA J. Ethics* **21**, 167–179 (2019).
36. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **169**, 866–872 (2018).
37. Inthout, J., Ioannidis, J. P. A., Borm, G. F. & Goeman, J. J. Small studies are more heterogeneous than large ones: A meta-meta-analysis. *J. Clin. Epidemiol.* **68**, 860–869 (2015).
38. Resche-Rigon, M., White, I. R., Bartlett, J. W., Peters, S. A. E. & Thompson, S. G. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Stat. Med.* **32**, 4890 (2013).
39. Chen, X. et al. Serological evidence of human infection with SARS-CoV-2: a systematic review and meta-analysis. *Lancet Glob. Health* **9**, e598 (2021).
40. Xie, Z., Ding, J., Jiao, J., Tang, S. & Huang, C. Screening instruments for early identification of unmet palliative care needs: a systematic review and meta-analysis. *BMJ Support Palliat. Care* **14**, 256 (2024).
41. Issitt, R. W. et al. Classification performance of neural networks versus logistic regression models: Evidence from healthcare practice. *Cureus* **14**, e22443 (2022).
42. Sculley, D. *et al.* Hidden technical debt in machine learning systems, in *Advances in Neural Information Processing Systems*, Vol. 2015-January (2015).
43. Mohammed, S. et al. The effects of data quality on machine learning performance on tabular data. *Inf. Syst.* **132**, 102549 (2025).

44. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-019-0048-x (2019).
45. Rudin, C. et al. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Stat. Surv.* **16**, 1–85 (2022).
46. Amann, J., Blasimme, A., Vayena, E., Frey, D. & Madai, V. I. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **20**, 310 (2020).
47. Wichmann, R. M. et al. Improving the performance of machine learning algorithms for health outcomes predictions in multicentric cohorts. *Sci. Rep.* **13**, 1022 (2023).
48. Ueda, D. et al. Fairness of artificial intelligence in healthcare: Review and recommendations. *Jpn. J. Radiol.* https://doi.org/10.1007/s11604-023-01474-3 (2024).
49. Banerjee, S., Lio, P., Jones, P. B. & Cardinal, R. N. A class-contrastive human-interpretable machine learning approach to predict mortality in severe mental illness. *NPJ Schizophr.* **7**, 60 (2021).
50. Meredith, J. W. et al. A comparison of the abilities of nine scoring algorithms in predicting mortality. *J. Trauma* **53**, 621 (2002).
51. Wang, Y. et al. A maintenance hemodialysis mortality prediction model based on anomaly detection using longitudinal hemodialysis data. *J. Biomed. Inform.* **123**, 103930 (2021).
52. Lin, S. Y. et al. Artificial intelligence prediction model for the cost and mortality of renal replacement therapy in aged and super-aged populations in Taiwan. *J. Clin. Med.* **8**, 995 (2019).
53. Liu, C. M. et al. Artificial intelligence-enabled model for early detection of left ventricular hypertrophy and mortality prediction in young to middle-aged adults. *Circ. Cardiovasc. Qual. Outcomes* **15**, e008360 (2022).
54. Shi, H. Y. et al. Artificial neural network model for predicting 5-year mortality after surgery for hepatocellular carcinoma: A nationwide study. *J. Gastrointest. Surg.* **16**, 2126 (2012).
55. Shi, L., Wang, X. C. & Wang, Y. S. Artificial neural network models for predicting 1-year mortality in elderly patients with intertrochanteric fractures in China. *Braz. J. Med. Biol. Res.* **46**, 993 (2013).
56. Puddu, P. E. & Menotti, A. Artificial neural networks versus proportional hazards Cox models to predict 45-year all-cause mortality in the Italian Rural Areas of the Seven Countries Study. *BMC Med. Res. Methodol.* **12**, 100 (2012).
57. Harris, A. H. S. et al. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty?. *Clin. Orthop. Relat. Res.* **477**, 452 (2019).
58. Jing, B. et al. Comparing machine learning to regression methods for mortality prediction using veterans affairs electronic health record clinical data. *Med. Care* **60**, 470 (2022).
59. Sakr, S. et al. Comparison of machine learning techniques to predict all-cause mortality using fitness data: The Henry Ford exercIse testing (FIT) project. *BMC Med. Inform. Decis. Mak.* **17**, 174 (2017).
60. Jones, B. E. et al. Computerized mortality prediction for community-acquired pneumonia at 117 veterans affairs medical centers. *Ann. Am. Thorac. Soc.* **18**, 1175 (2021).
61. Singh, A. et al. Deep learning for explainable estimation of mortality risk from myocardial positron emission tomography images. *Circ. Cardiovasc. Imaging* **15**, e014526 (2022).
62. Lu, M. T. et al. Deep learning to assess long-term mortality from chest radiographs. *JAMA Netw .Open* **2**, e197416 (2019).
63. Ulloa Cerna, A. E. et al. Deep-learning-assisted analysis of echocardiographic videos improves predictions of all-cause mortality. *Nat. Biomed. Eng.* **5**, 546 (2021).
64. Wang, L. et al. Development and validation of a deep learning algorithm for mortality prediction in selecting patients with dementia for earlier palliative care interventions. *JAMA Netw. Open* **2**, e196972 (2019).
65. Mohammad, M. A. et al. Development and validation of an artificial neural network algorithm to predict mortality and admission to hospital for heart failure after myocardial infarction: A nationwide population-based study. *Lancet Digit. Health* **4**, e37 (2022).
66. Valsaraj, A. et al. Development and validation of echocardiography-based machine-learning models to predict mortality. *EBioMedicine* **90**, 104479 (2023).
67. Li, Z. et al. Development and validation of questionnaire-based machine learning models for predicting all-cause mortality in a representative population of China. *Front. Public Health* **11**, 1033070 (2023).
68. Zhou, J. et al. Development of an electronic frailty index for predicting mortality and complications analysis in pulmonary hypertension using random survival forest model. *Front. Cardiovasc. Med.* **9**, 735906 (2022).
69. Giang, K. W., Helgadottir, S., Dellborg, M., Volpe, G. & Mandalenakis, Z. Enhanced prediction of atrial fibrillation and mortality among patients with congenital heart disease using nationwide register-based medical hospital data and neural networks. *Eur. Heart J. Digit. Health* **2**, 568 (2021).
70. Castela Forte, J. et al. Ensemble machine learning prediction and variable importance analysis of 5-year mortality after cardiac valve and CABG operations. *Sci. Rep.* **11**, 3467 (2021).
71. Hernesniemi, J. A. et al. Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome–the MADDEC study. *Ann. Med.* **51**, 156 (2019).
72. Qiu, W. et al. Interpretable machine learning prediction of all-cause mortality. *Commun. Med.* **2**, 125 (2022).
73. Niedziela, J. T. et al. Is neural network better than logistic regression in death prediction in patients after ST-segment elevation myocardial infarction?. *Kardiol. Pol.* **79**, 1353 (2021).
74. Mostafaei, S. et al. Machine learning algorithms for identifying predictive variables of mortality risk following dementia diagnosis: A longitudinal cohort study. *Sci. Rep.* **13**, 9480 (2023).
75. Cui, Y. et al. Machine learning approaches for prediction of early death among lung cancer patients with bone metastases using routine clinical characteristics: An analysis of 19,887 patients. *Front. Public Health* **10**, 1019168 (2022).
76. Parikh, R. B. et al. Machine learning approaches to predict 6-month mortality among patients with cancer. *JAMA Netw. Open* **2**, 1019168 (2019).
77. Tong, J. et al. Machine learning can predict total death after radiofrequency ablation in liver cancer patients. *Clin. Med. Insights Oncol.* **15**, 11795549211000016 (2021).
78. de Capretz, P. O. et al. Machine learning for early prediction of acute myocardial infarction or death in acute chest pain patients using electrocardiogram and blood tests at presentation. *BMC Med. Inform. Decis. Mak.* **23**, 25 (2023).
79. Tian, P. et al. Machine learning for mortality prediction in patients with heart failure with mildly reduced ejection fraction. *J. Am. Heart Assoc.* **12**, e029124 (2023).
80. Mamprin, M. et al. Machine learning for predicting mortality in transcatheter aortic valve implantation: An inter-center cross validation study. *J. Cardiovasc. Dev. Dis.* **8**, 65 (2021).
81. Motwani, M. et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: A 5-year multicentre prospective registry analysis. *Eur. Heart J.* **38**, 500–507 (2017).
82. dos Santos, H. G., do Nascimento, C. F., Izbicki, R., de Duarte, Y. A. O. & Filho, A. D. P. C. Machine learning for predictive analyses in health: An example of an application to predict death in the elderly in São Paulo, Brazil. *Cad Saude Publica* **35**, e00050818 (2019).
83. Feng, X. et al. Machine learning improves mortality prediction in three-vessel disease. *Atherosclerosis* **367**, 1–7 (2023).
84. Yu, Y. et al. Machine learning methods for predicting long-term mortality in patients after cardiac surgery. *Front. Cardiovasc. Med.* **9**, 831390 (2022).
85. Tamminen, J., Kallonen, A., Hoppu, S. & Kalliomäki, J. Machine learning model predicts short-term mortality among prehospital patients: A prospective development study from Finland. *Resusc. Plus* **5**, 100089 (2021).

86. Xu, C., Subbiah, I. M., Lu, S. C., Pfob, A. & Sidey-Gibbons, C. Machine learning models for 180-day mortality prediction of patients with advanced cancer using patient-reported symptom data. *Qual. Life Res.* **32**, 713 (2023).
87. Katsiferis, A. et al. Machine learning models of healthcare expenditures predicting mortality: A cohort study of spousal bereaved Danish individuals. *PLoS ONE* **18**, e0289632 (2023).
88. Kanda, E. et al. Machine learning models predicting cardiovascular and renal outcomes and mortality in patients with hyperkalemia. *Nutrients* **14**, 4614 (2022).
89. Lu, J. et al. Machine learning risk prediction model for acute coronary syndrome and death from use of non-steroidal anti-inflammatory drugs in administrative data. *Sci. Rep.* **11**, 18314 (2021).
90. Scrutinio, D. et al. Machine learning to predict mortality after rehabilitation among patients with severe stroke. *Sci. Rep.* **10**, 20127 (2020).
91. Li, Y. M. et al. Machine learning to predict the 1-year mortality rate after acute anterior myocardial infarction in Chinese patients. *Ther. Clin. Risk Manag.* **16**, 1–16 (2020).
92. Guo, R. et al. Machine learning-based approaches for prediction of patients' functional outcome and mortality after spontaneous intracerebral hemorrhage. *J. Pers. Med.* **12**, 112 (2022).
93. Li, Y. et al. Machine learning-based models to predict one-year mortality among Chinese older patients with coronary artery disease combined with impaired glucose tolerance or diabetes mellitus. *Cardiovasc. Diabetol.* **22**, 139 (2023).
94. Asrian, G., Suri, A. & Rajapakse, C. Machine learning-based mortality prediction in hip fracture patients using biomarkers. *J. Orthop. Res.* **42**, 395 (2024).
95. Ivanics, T. et al. Machine learning–based mortality prediction models using national liver transplantation registries are feasible but have limited utility across countries. *Am. J. Transpl.* **23**, 64 (2023).
96. Behnoush, A. H. et al. Machine learning-based prediction of 1-year mortality in hypertensive patients undergoing coronary revascularization surgery. *Clin. Cardiol.* **46**, 269 (2023).
97. Kampaktsis, P. N. et al. Machine learning-based prediction of mortality after heart transplantation in adults with congenital heart disease: A UNOS database analysis. *Clin. Transpl.* **37**, e14845 (2023).
98. Lin, Y. J. et al. Machine-learning monitoring system for predicting mortality among patients with noncancer end-stage liver disease: Retrospective study. *JMIR Med. Inform.* **8**, e24305 (2020).
99. Lin, F. Y. et al. Mortality impact of low CAC density predominantly occurs in early atherosclerosis: Explainable ML in the CAC consortium. *J. Cardiovasc. Comput. Tomogr.* **17**, 28 (2023).
100. Liu, Y. et al. Nomogram and machine learning models predict 1-year mortality risk in patients with sepsis-induced cardiorenal syndrome. *Front. Med. (Lausanne)* **9**, 792238 (2022).
101. Forssten, M. P., Bass, G. A., Ismail, A. M., Mohseni, S. & Cao, Y. Predicting 1-year mortality after hip fracture surgery: An evaluation of multiple machine learning approaches. *J. Pers. Med.* **11**, 727 (2021).
102. Alimbayev, A. et al. Predicting 1-year mortality of patients with diabetes mellitus in Kazakhstan based on administrative health data using machine learning. *Sci. Rep.* **13**, 8412 (2023).
103. El-Bouri, W. K., Sanders, A. & Lip, G. Y. H. Predicting acute and long-term mortality in a cohort of pulmonary embolism patients using machine learning. *Eur. J. Intern. Med.* **118**, 42 (2023).
104. Park, J. et al. Predicting long-term mortality in patients with acute heart failure by using machine learning. *J. Card Fail.* **28**, 1078 (2022).
105. Penso, M. et al. Predicting long-term mortality in TAVI patients using machine learning techniques. *J. Cardiovasc. Dev. Dis.* **8**, 44 (2021).
106. Guo, A., Mazumder, N. R., Ladner, D. P. & Foraker, R. E. Predicting mortality among patients with liver cirrhosis in electronic health records with machine learning. *PLoS ONE* **16**, e0256428 (2021).
107. Abedi, V. et al. Predicting short and long-term mortality after acute ischemic stroke using EHR. *J. Neurol. Sci.* **427**, 117560 (2021).
108. Zhou, J. et al. Predicting stroke and mortality in mitral regurgitation: A machine learning approach. *Curr. Probl. Cardiol.* https://doi.org/10.1016/j.cpcardiol.2022.101464 (2023).
109. Rauf, A. et al. Predicting stroke and mortality in mitral stenosis with atrial flutter: A machine learning approach. *Ann. Noninvasive Electrocardiol.* **28**, e13078 (2023).
110. Shi, N. et al. Predicting the need for therapeutic intervention and mortality in acute pancreatitis: A two-center international study using machine learning. *J. Pers. Med.* **12**, 616 (2022).
111. Zhou, Y. et al. Prediction of 1-year mortality after heart transplantation using machine learning approaches: A single-center study from China. *Int. J. Cardiol.* **339**, 21 (2021).
112. Lee, H. C. et al. Prediction of 1-year mortality from acute myocardial infarction using machine learning. *Am. J. Cardiol.* **133**, 23 (2020).
113. Tran, N. T. D. et al. Prediction of all-cause mortality for chronic kidney disease patients using four models of machine learning. *Nephrol. Dial. Transpl.* **38**, 1691 (2023).
114. Raghunath, S. et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat. Med.* **26**, 886 (2020).
115. Kawano, K. et al. Prediction of mortality risk of health checkup participants using machine learning-based models: The J-SHC study. *Sci. Rep.* **12**, 141546 (2022).
116. Weng, S. F., Vaz, L., Qureshi, N. & Kai, J. Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLoS ONE* **14**, e0214365 (2019).
117. Zhou, Q. et al. Prediction of premature all-cause mortality in patients receiving peritoneal dialysis using modified artificial neural networks. *Aging* **13**, 14170 (2021).
118. Huang, S. H., Loh, J. K., Tsai, J. T., Houg, M. F. & Shi, H. Y. Predictive model for 5-year mortality after breast cancer surgery in Taiwan residents. *Chin. J. Cancer* **36**, 1–9 (2017).
119. Sheng, K. et al. Prognostic machine learning models for first-year mortality in incident hemodialysis patients: Development and validation study. *JMIR Med. Inform.* **8**, e20578 (2020).
120. Zachariah, F. J., Rossi, L. A., Roberts, L. M. & Bosserman, L. D. Prospective comparison of medical oncologists and a machine learning model to predict 3-month mortality in patients with metastatic solid tumors. *JAMA Netw. Open* https://doi.org/10.1001/jamanetworkopen.2022.14514 (2022).
121. Unterhuber, M. et al. Proteomics-enabled deep learning machine algorithms can enhance prediction of mortality. *J. Am. Coll. Cardiol.* **78**, 1621 (2021).
122. Wu, X. D. et al. Risk factors prediction of 6-month mortality after noncardiac surgery of older patients in China: A multicentre retrospective cohort study. *Int. J. Surg.* **110**, 219 (2024).
123. Hwangbo, L. et al. Stacking ensemble learning model to predict 6-month mortality in ischemic stroke patients. *Sci. Rep.* **12**, 17389 (2022).
124. Ross, E. G. et al. The use of machine learning for the identification of peripheral artery disease and future mortality risk. *J. Vasc. Surg.* **64**, 1515 (2016).
125. Wang, H. *et al.* Using machine learning to integrate socio-behavioral factors in predicting cardiovascular-related mortality risk, in *Studies in Health Technology and Informatics*, Vol. 264 (2019).

Acknowledgements
Delpino FM received a post-doctoral fellowship from the National Council for Scientific and Technological Development (CNPq), Call 07/2022, in conjunction with FAPERGS—Foundation for Research Support of the State of Rio Grande do Sul (FAPERGS), during the writing of the manuscript. Nunes BP and Chiavegatto Filho AD received a research productivity grant from CNPq. Alexandre.

Author contributions
Delpino FM participated in the design, writing, and leadership of the article. Nunes BP and Chiavegatto Filho AD contributed to the conception and development of the idea, design, and critical review of the paper. Pimenta LP, Gonzales DF, Victor A, Araújo C, and Moura KA participated in the selection of studies and data extraction. Miranda JJ and Batista SRR contributed to the interpretation of the data and critical review of the paper. All authors reviewed and contributed significantly to the writing of the manuscript.

Funding
National Council for Scientific and Technological Development (CNPq), Call 07/2022, in conjunction with FAPERGS—Foundation for Research Support of the State of Rio Grande do Sul (FAPERGS), during the writing of the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-26714-6.

Correspondence and requests for materials should be addressed to F.M.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025