Machine-Learning based Prediction Models for Healthcare Outcomes in Patients Participating in Cardiac Rehabilitation: A Systematic Review

Xiarepati Tieliwaerdi¹; Kathryn Manalo¹; Abulikemu Abuduweili²; Sana Khan¹; Edmund Appiah-kubi; Andrew Oehler³

1: Department of Medicine, Allegheny Health Network, Pittsburgh, Pennsylvania
2. Institute of Robotics, Carnegie Mellon University, Pittsburgh, Pennsylvania
3. Allegheny Health Network Cardiovascular Institute, Allegheny Health Network, Pittsburgh, Pennsylvania

Disclosure

We confirm that all authors have read and approved the submission of this manuscript.

This manuscript has no relationship with industry, and no competing interests exist.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

The work was not funded.
Abstract

Purpose: CR has been proven to reduce mortality and morbidity in patients with CVD. ML techniques are increasingly used to predict healthcare outcomes in various fields of medicine including CR. This systemic review aims to perform critical appraisal of existing ML based prognosis predictive model within CR and identify key research gaps in this area.

Review methods: A systematic literature search was conducted in Scopus, PubMed, Web of Science and Google Scholar from the inception of each database to 28th January 2024. The data extracted included clinical features, predicted outcomes, model development and validation as well as model performance metrics. Included studies underwent quality assessments using the IJMEDI.

Summary: 22 ML-based clinical models from 7 studies across multiple phases of CR were included. Most models were developed using smaller patient cohorts from 41 to 227, with one exception involving 2280 patients. The prediction objectives ranged from patient intention to initiate CR to graduate from outpatient CR along with interval physiological and psychological response to CR. The best-performing ML models reported AUC between 0.82 and 0.91, sensitivity from 0.77 to 0.95, indicating good prediction capabilities. However, none of them underwent calibration or external validation. Most studies raised concerns for bias. Readiness of these models for implement into practice is questionable. External validation of existing models and development of new models with robust methodology based on larger populations and targeting diverse clinical overcomes in CR are needed.

Introduction

Cardiovascular disease (CVD) remains a leading cause of morbidity and mortality globally, causing over 19 million deaths alone in 2020 worldwide\(^1\). Cardiac rehabilitation (CR) is a comprehensive evidenced-based intervention tailored to patients with CVD conditions like ischemic heart disease, heart failure, myocardial infarction (MI), or those undergoing cardiovascular interventions such as coronary angioplasty or bypass grafting\(^2-5\). It encompasses exercise training, health behavior modification, patient education, nutritional and psychological counseling, proven to reduce morbidity and mortality, enhance functional capacity, and improve quality of life in CVD patients\(^6-11\). CR comprises of three phases in general. Phase I occurs post-cardiovascular events during the acute inpatient setting for stabilization and minimizing deconditioning. Phase II begins once the patient is able to be safely discharged from the hospital. A patient can expect to participate in outpatient CR focusing on supervised exercise program and medical education generally up to 12 weeks with 36 sessions in total. After completion, patients enter the phase III as a long-term maintenance period, where patients continue their rehab exercises and lifestyle changes on their own at home or within a community setting\(^12-14\).
CR is a complex medical entity, where challenges lie in multiple aspects including unsatisfactory patient enrollment and adherence, varying clinical responses and the need for ongoing treatment adjustment to improve ultimate outcomes\textsuperscript{15–18}. CR program involves a large amount of multimodal electronic health records (EHRs), which contain a wide range of variables in diverse data formats such as demographic information, providers' free-text notes, laboratory and imaging findings. These data are often too voluminous and technically unfeasible to process by traditional analytical methods\textsuperscript{19}. In addition, traditional statistical methods are built on specific assumptions, including specific error distributions, the additive linear predictor's parameters, and the proportionality of hazards\textsuperscript{20}. However, these assumptions do not always align with the realities encountered in clinical settings. Moreover, traditional methods primarily aim to infer relationships between variables, particularly interactions between a main determinant and individual confounders, they are insufficient in current clinical settings, where there are a growing number of variables and an increasing need for early intervention through outcome prediction\textsuperscript{21}. The advent of machine learning (ML) in healthcare offers a potential solution to these challenges. ML detects intricate interactions among millions of complex variables without being programmed with hypophysis\textsuperscript{22}. It is particularly well-suited for analyzing large datasets with numerous features and high complexities such as EHR. It can analyze various data types and is better at capturing non-linear relationships between variables\textsuperscript{20,23}. ML based predictive models have been widely proposed in CVD to aid decision-making processes for clinicians, showing promise in predicting patient specific clinical progression and health care outcomes\textsuperscript{24–26}.

Recent literature has seen a proliferation of systematic reviews and meta-analyses targeted at evaluating the performance of ML-based clinical prediction models in different fields of medicine\textsuperscript{27–29}. However, to the best of our knowledge, there has yet to be a systematic review of ML predictive models in CR. The primary aim of this review is to evaluate existing literature regarding ML predictive model in the context of CR and provide systemic appraisal of these models from their development to validation. We aim to compare performance metrics, validation processes and appropriateness of algorithm used in CR, as well as identify key research gaps in this area.

**Methods**

**Literature Search Strategy**

The literature search and related statistical analyses were conducted in accordance with the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement\textsuperscript{30}. We comprehensively searched publications from database inception to 28th January 2024 in Google Scholar, PubMed, Web of Science, and Scopus. Keywords used for the search included a combination of terms related to ML and CR. The full search strategy, including the specific combinations of search terms used in each database, was provided in supplementary table 1.
Inclusion and Exclusion Criteria

In our review, we included studies that employed a ML model to predict clinical progression, health outcomes, or risk stratification in a cohort of adult patients participating in any phase of CR. We did not restrict it by the country of origin or publication source. For clarity, we defined ML as algorithms, such as random forest (RF) analysis, support vector machines (SVM), and neural networks, that are more complex than logistic regression models and capable of making decisions based on data patterns.

Our inclusion criteria were structured using the population, intervention, comparison, outcome, and time (PICOT) approach as follows\textsuperscript{31}: 1) Population: Adult patients aged 18 years or older engaging in CR programs; 2) Intervention: Studies used ML models for predicting outcomes in CR; 3) Comparison: Not applicable, due to the lack of a universally accepted prognostic model in CR; 4) Outcome: Studies reporting on clinical progression, health outcomes or risk stratification outcomes; 5) Time: Studies which harnessed features to predict outcomes after any given follow-up period is accepted. In terms of study design, we did not restrict the types of studies included. Retrospective, prospective, cross-sectional, cohort studies, case-control studies and randomized controlled trials were all considered for inclusion.

Exclusion criteria were: 1) Studies not in English; 2) Studies not involving ML-based predictive models: for example, studies reporting novel ML based wearable devices used for monitoring biomarkers such as heart rate or blood pressure were excluded; 3) Studies where predictive models are not developed from patients in CR; 4) Studies focusing on identifying predictors associated with outcomes rather than develop a prognostic model. 5) Reviews, case reports and studies not available in full text were excluded.

Data Extraction

Duplicate records were initially removed using auto-deduplication function in EndNote 21, followed by a manual check for complete deduplication. The screening of titles and abstracts was then carried out in EndNote 21, adhering to the inclusion and exclusion criteria previously outlined.

A team of four reviewers (XT, KM, SK and EA) assessed articles for eligibility first by screening titles and abstracts to ensure relevance in EndNote; each study was independently assessed by at least two reviewers. Both agreed-upon and conflicting articles were retained for second-round screening based on full-text review. The full-text evaluation was independently carried out by the reviewers (XT and AA). Disagreements were resolved through mutual consensus.

The data from included articles were extracted into tables by two authors (SK and KM) independently. Cross-check was conducted. Variables were extracted and tabulated in Excel 2020, which includes 1) baseline characteristics of studies; 2) features and ML algorithms used for the development of the models; 3) objectives of the prediction models as well as evaluation index for their performance, including area under the receiver operating characteristic curve (AUROC), accuracy, precision, sensitivity and specificity; 4) methods employed for model validation.
Quality Assessment

The quality of the included studies was evaluated independently by two reviewers (XT and AA) by using IJMEDI checklist\(^{32}\). The IJMEDI checklist is developed specifically for use in medical ML studies to distinguish high-quality ML research from studies with mere application of ML methods to medical data. It focuses on six key aspects: problem and data understanding, data preparation, modeling, validation, and deployment, encompassing a total of 30 detailed questions. Responses to these questions are categorized as OK (adequately addressed), mR (moderately addressed with improvement potential), or MR (inadequately addressed). For questions deemed high priority, scores were assigned as 0 for OK, 1 for mR, and 2 for MR. For low-priority questions, the scoring was assigned as 0 for OK, 1 for mR, and 2 for MR. Studies were classified into three quality categories: low (0–19.5 points), medium (20–34.5 points), or high quality (35–50 points).

Results

From the initial search that yielded 151 records, 81 were excluded following a review of the title and abstract, 60 records were excluded after full-text review. In the end, 7 studies met the inclusion criteria and were included in our systemic review (Figure 1 shows process of screening and study selection)\(^{35–41}\).

Figure 1: PRISMA flowchart of the review process
Study characteristics

Table 1 summarizes characteristics of studies using ML prediction model in CR. Seven studies are included, one of which is a multicenter study conducted in Belgium and Ireland\textsuperscript{37}. The rest are single-center studies from Australia\textsuperscript{35,40}, Belgium\textsuperscript{38}, Italy\textsuperscript{41}, Malaysia\textsuperscript{36}, and Chile\textsuperscript{39}. Study designs include five retrospective studies\textsuperscript{35–38,41}, one cross-sectional study\textsuperscript{40}, and one that combines retrospective and prospective approaches\textsuperscript{39}. Most studies target phase II of CR, with one addressing phases II and III\textsuperscript{35} and another focusing solely on phase III\textsuperscript{37}. Participant ages were reported in four of the seven studies\textsuperscript{35,38,40,41}, with mean ages ranging from 63 to 67.98 years. The percentage of female participants is reported in three studies and varies from 18.60\% to 27.00\%\textsuperscript{38,40,41}, while the other four studies do not report gender distribution. Patients included were typically referred to CR, with one study specifying the prerequisite of being employed prior to a cardiac event\textsuperscript{36}. The studies generally excluded patients lacking post-rehab data, those with contraindications to exercise, or those unwilling to participate. Functional sample sizes range significantly from 41 to 2280 across the studies.

<table>
<thead>
<tr>
<th>Study (year)</th>
<th>Study location</th>
<th>Rehab phase</th>
<th>Study design</th>
<th>Age (years old)</th>
<th>Female (%)</th>
<th>Cohort inclusion criteria</th>
<th>Cohort exclusion criteria</th>
<th>Functional sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Van et al (2010)\textsuperscript{35}</td>
<td>Australia</td>
<td>II and III</td>
<td>Retrospective</td>
<td>Mean: 65.35</td>
<td>NR</td>
<td>Patients referred to CR</td>
<td>Patients did not have post-rehab data or where values in target attributes were missing.</td>
<td>2280</td>
</tr>
<tr>
<td>Lofaro et al (2016) \textsuperscript{41}</td>
<td>Italy</td>
<td>II</td>
<td>Retrospective</td>
<td>Mean: 67.98</td>
<td>18.60</td>
<td>Patients admitted for CR after CAD or MI</td>
<td>NR</td>
<td>129</td>
</tr>
<tr>
<td>De Cannière et al (2020) \textsuperscript{38}</td>
<td>Belgium</td>
<td>II</td>
<td>Retrospective</td>
<td>Mean: 63</td>
<td>27.00</td>
<td>Patients eligible for CR</td>
<td>Patients who were unable to exercise or didn't complete the total study.</td>
<td>89</td>
</tr>
<tr>
<td>Jahandideh et al (2021) \textsuperscript{40}</td>
<td>Australia</td>
<td>II</td>
<td>Cross-sectional</td>
<td>Mean: 63</td>
<td>26.27</td>
<td>Patients referred to outpatient CR</td>
<td>Patients who were too ill or unable to speak English.</td>
<td>217</td>
</tr>
<tr>
<td>Study</td>
<td>Country</td>
<td>Study Type</td>
<td>Age</td>
<td>Follow-up</td>
<td>Referral Criteria</td>
<td>Other Criteria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------------</td>
<td>------------</td>
<td>---------</td>
<td>-----------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yuan et al (2023) 36</td>
<td>Malaysia</td>
<td>II</td>
<td>≥18</td>
<td>NR</td>
<td>Patients referred to CR and employed prior to having a cardiac event.</td>
<td>Patients who were pensioners or unemployed.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filos et al (2023) 37</td>
<td>Belgium and Ireland</td>
<td>III</td>
<td>40-80</td>
<td>NR</td>
<td>Patients completed a phase II CR and enrolled in Internet based home CR</td>
<td>Patients who dropped out or exercised very sparsely.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torres et al (2023) 39</td>
<td>Chile</td>
<td>II</td>
<td>≥18</td>
<td>NR</td>
<td>Patients referred to CR.</td>
<td>Patients who had any contraindication to physical exercise.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: CAD: coronary artery disease

Table 1. Characteristics of studies using ML prediction model in CR

Model characteristics

Table 2 presented a summary of characteristics of ML model characteristics included in our review. Each study has different prediction goals, ranging from forecasting a patient’s intention to begin CR to estimating their adherence to the program. Two studies aimed to predict patients’ prognosis in CR including 6-minute walk distance38, various physiological and psychological outcome35. Additionally, two studies focused on predicting patient disposition post-CR, such as likelihood of returning to work36 as well as the transition from one phase of CR to another39.

These studies utilized a variety of patient features to develop their models. These features encompassed anthropometric measurements such as body mass index (BMI) and waist circumference, demographic information and medical history especially cardiovascular health. Psycho-behavioral profiles were also considered in most of the studies including evaluation of anxiety and depression levels. Laboratory test results and physical fitness levels were commonly used. Two studies harnessed specific imaging like EKG and TTE38,39.

The number of features used across the studies varies, with one utilizing as few as 17 features41 and others up to 8240. Most studies used pre-processing techniques for feature selection to reduce the number of input variables to those that are most important to the predictive model to improve performance and reduce computational cost. RF and principal component analysis (PCA) were most employed35,39,40.

Only two of the seven studies reported methods to handle missing data35,41. In terms of ML algorithms, there was a wide array employed across these studies, from basic decision trees (DT) and SVM to advanced ensemble methods like AdaBoost and XGBoost. Three of seven studies
used a single algorithm\textsuperscript{35,38,40}, the other four compared multiple different algorithms, with DT, RF, and SVM being the most adopted\textsuperscript{36,37,39,41}.

<table>
<thead>
<tr>
<th>Study (year)</th>
<th>Object of prediction</th>
<th>Features used for prediction</th>
<th>Missing data handling</th>
<th>Feature selection</th>
<th>ML algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Van et al (2010)\textsuperscript{38}</td>
<td>Physiological and psychosocial outcomes of CR</td>
<td>Anthropometric measurements, sociodemographic, psycho-behavioral profile, medical hx, laboratory tests, physical fitness level</td>
<td>49 Cases with &gt;25% of missing data were discarded, otherwise were replaced by mean or mode</td>
<td>RF, PCA</td>
<td>DT</td>
</tr>
<tr>
<td>Lofaro et al (2016)\textsuperscript{41}</td>
<td>Patient-specific CR exercise program</td>
<td>Anthropometric measurements, sociodemographics, medical hx, laboratory tests and images (PFTs), physical fitness level</td>
<td>17 Variables with &gt;50% of missing data were discarded.</td>
<td>None</td>
<td>Lasso regression, SVM, RF, Bagged FDA, Boost C 5.0, Bagged CART</td>
</tr>
<tr>
<td>De Cannière et al (2020)\textsuperscript{38}</td>
<td>6-minute walk distance</td>
<td>Anthropometric measurements, psycho-behavioral profile, medical hx, laboratory tests and images (EKG, TTE), physical fitness level</td>
<td>NR</td>
<td>NR</td>
<td>SVM, linear regression</td>
</tr>
<tr>
<td>Jahandideh et al (2021)\textsuperscript{40}</td>
<td>Individual's intention to engage in CR</td>
<td>Anthropometric measurements, socio-demographics, medical hx, perceived need for CR</td>
<td>82</td>
<td>NR</td>
<td>RF</td>
</tr>
<tr>
<td>Yuan et al (2023)\textsuperscript{36}</td>
<td>Ability to return to work after CR.</td>
<td>Anthropometric measurements, socio-demographic and psychosocial profile, medical hx, laboratory tests, physical fitness level</td>
<td>30</td>
<td>NR</td>
<td>Recursive feature elimination</td>
</tr>
<tr>
<td>Filos et al (2023)\textsuperscript{37}</td>
<td>6-month adherence to home-based CR</td>
<td>Anthropometric measurements, psycho-behavioral profile, laboratory tests, physical fitness level</td>
<td>52</td>
<td>NR</td>
<td>DT and RF</td>
</tr>
<tr>
<td>Torres et al (2023)\textsuperscript{39}</td>
<td>Probability of progression from CR phase II to III</td>
<td>Anthropometric measurements, psycho-behavioral profile, laboratory tests and images (EKG and TTE), physical fitness level</td>
<td>44</td>
<td>NR</td>
<td>XGboost, gradient boosting, SVM, RF, KNN</td>
</tr>
</tbody>
</table>
Table 2. Characteristics of ML prediction model

Performance and validation

Table 3 outlined the best ML model with its performance metrics and validation approaches used to evaluate the ML models. The models’ performance was evaluated against various metrics such as area under the curve (AUC), sensitivity and specificity, mean absolute error (MAE), and normalized mean squared error (NMSE). The selection of best-performing ML algorithm differed per study. Six out of seven employed internal validation techniques, with cross-validation being the most prevalent method\textsuperscript{35–39,41}. None of the studies underwent external validation, meaning they were not validated by using data from populations different from those used to develop the models, nor were they implemented in real-life practice.

<table>
<thead>
<tr>
<th>Study (year)</th>
<th>Best ML algorithm</th>
<th>Performance metrics</th>
<th>Validation methods</th>
<th>External validation</th>
<th>Practice implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Van et al (2010)\textsuperscript{35}</td>
<td>DT</td>
<td>AUC: 0.815; accuracy: 0.769; precision: 0.734; sensitivity: 0.769</td>
<td>Internal validation: 10-fold cross validation</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Lofaro et al (2016)\textsuperscript{41}</td>
<td>Lasso Regression</td>
<td>Accuracy: 0.935; precision: 0.941; sensitivity: 0.9</td>
<td>5-fold cross-validation;</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>De Cannière et al (2020)\textsuperscript{38}</td>
<td>SVM</td>
<td>MAE: 42.8 m</td>
<td>20-fold-validation;</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Jahandideh et al (2021)\textsuperscript{40}</td>
<td>RF</td>
<td>Accuracy: 0.715</td>
<td>NR</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Yuan et al (2023)\textsuperscript{36}</td>
<td>AdaBoost</td>
<td>AUC: 0.924; accuracy: 0.864; sensitivity: 0.928; specificity: 0.733</td>
<td>10-fold cross validation</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Filos et al (2023)\textsuperscript{37}</td>
<td>DT</td>
<td>Sensitivity: 0.945; precision: 0.80</td>
<td>10-fold cross validation</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Torres et al (2023)\textsuperscript{39}</td>
<td>XGboost</td>
<td>NMSE: 0.008; $R^2$: 92%</td>
<td>10-fold cross validation</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>

Abbreviations: DT: decision tree; RF: random forest; AUC: area under curve; SVM: support vector machine. MAE: mean absolute error (MAE); NMSE: normalized mean squared error (NMSE); $R^2$: R-squared.
Table 4 summarized the scores for each dimension and the total score of each study according to the IMED checklist. The included studies had an average score of 30.8, with a range from 26 to 35.5. Most of the studies fell into the medium quality category, one study stood out as being of high quality\(^{35}\). Most studies demonstrated a discernible bias in the 'Data Preparation', 'Validation', and 'Deployment' dimensions, with lower scores that suggest these areas may have impacted the overall quality and reliability of the outcomes.

<table>
<thead>
<tr>
<th>Study(year)</th>
<th>Problem Understanding (10)</th>
<th>Data Understanding (6)</th>
<th>Data Preparation (8)</th>
<th>Modeling (6)</th>
<th>Validation (12)</th>
<th>Deployment (8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Van et al (2010)</td>
<td>8.5</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>2.5</td>
</tr>
<tr>
<td>Lofaro et al (2016)</td>
<td>8.5</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>1.5</td>
</tr>
<tr>
<td>De Cannière et al (2020)</td>
<td>9.5</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Jahandideh et al (2021)</td>
<td>9</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Yuan et al (2023)</td>
<td>9.5</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>8.5</td>
<td>3</td>
</tr>
<tr>
<td>Filos et al (2023)</td>
<td>9.5</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>8.5</td>
<td>3</td>
</tr>
<tr>
<td>Torres et al (2023)</td>
<td>9</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>8.5</td>
<td>3</td>
</tr>
</tbody>
</table>
Table 4 Results of the quality assessment according to the IJMEDI checklist

<table>
<thead>
<tr>
<th></th>
<th>0.5</th>
<th>0.5</th>
<th>0.5</th>
<th>1</th>
<th>1</th>
<th>0.5</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total (50)</td>
<td>35</td>
<td>26</td>
<td>30</td>
<td>30</td>
<td>33</td>
<td>29</td>
<td>33.5</td>
</tr>
</tbody>
</table>

*Items in bold indicate high priority. 2, 1, 0 points for high-priority requirements that are, respectively, OK, mR and MR; assign half these scores (i.e., 1, 0.5 and 0) for low-priority requirement.

Discussion

This review has evaluated the performance of 22 ML based clinical models in 7 studies aiming to predict healthcare outcomes for patients participating in CR. The prediction objectives ranged from patient intention to initiate CR to graduate from outpatient CR along with interval physiological and psychological changes during the program. The best-performing ML models in their respective tasks reported AUC between 0.82 and 0.91, sensitivity from 0.77 to 0.95; demonstrating good prediction capabilities in general. The majority of the included studies were rated as medium quality according to the IJMEDI checklist and there were high concerns for bias. Meta-analysis was not conducted as the included ML models were highly heterogenous in terms of targeted population, prediction objectives, outcome measurement and validation.

An ideal clinical prediction model should correctly distinguish between patients who will develop certain events and those who will not without misclassification in any case. Its quality is associated with two properties of the model: discrimination and calibration. Discrimination is the model's capacity to correctly separate individuals at higher risk of an event from those at lower risk. Calibration refers to the model’s ability to estimate absolute risks accurately. Discrimination is typically measured by AUC of receiver operating characteristic (ROC) curve, it can also be assessed by sensitivity and specificity. However, sensitivity and specificity vary as the cut point used to determine “positive” and “negative” test results change. The ROC curve is a graph of the sensitivity of a test versus its false-positive rate (1-specificity) for all potential cut points. The AUC-ROC represents average prediction accuracy after balancing the inherent tradeoffs that exist between sensitivity and specificity across the spectrum of varying cut points. A higher AUC indicates better discrimination ability. An AUC of 0.5 suggests no discrimination, equivalent to random guessing, while an AUC of 1.0 indicates perfect discrimination. Van et al reported an AUC of 0.815, which suggests good discrimination in predicting post-rehab deterioration. Yuan et al. reported an AUC of 0.923, reflecting the model's excellent effectiveness in predicting patients’ likelihood of patients’ return-to-work. Two studies reported sensitivity and precision rather than AUC as performance metrics, reflecting model’s performance for a particular cut point instead of all possible threshold. One study reported accuracy only, it measures the proportion...
of true results, both true positives and true negatives. High accuracy can sometimes be misleading if the class distribution is uneven. Jahandideh et al. claimed an accuracy of 0.715 in differentiating highly motivated patients for CR initiation but did not report distribution of motivation levels in the study population. This could lead to overestimation of accuracy in a predominantly motivated group or underestimation in a less motivated one. Overall, the reviewed studies used various evaluation metrics, appropriateness of which are generally acceptable, but could benefit from a more comprehensive measurement approach.

Model calibration is typically assessed after its discrimination is deemed acceptable. Calibration reflects the concordance between model prediction and observed outcomes. It is often represented by calibration plots or tables comparing predicted and observed risk. None of the included ML models underwent calibrations, so their ability to predict absolute risk remains uncertain. Although they have demonstrated overall good discrimination, calibration is essential to prove their capability in clinical decision support. This limitation contributed to lower scores in the deployment section of the IJMEDI checklist in table 4. Another major deficiency is the lack of external validation, which tests a model's efficacy in a different population than it was initially derived from. Without robust external validation, a model's generalizability is questionable. One notable example is that National Institute for Health and Care Excellence recommended an independent external validation of QRISK2 and the Framingham risk score, which were performed subsequently and demonstrated systematic miscalibration of Framingham risk scores and led to the need for different treatment thresholds in UK cohort. Although some cases may not require immediate external validation if the sample size is large and representative of predictors and target outcomes, most studies in our review, with populations ranging from 41 to 227 (except one with 2280 patients), would benefit from external validation as a key step towards implementation into clinical practice.

In addition to methodological robustness in model development, a useful clinical prediction model should address clinically significant issues. Low enrollment and adherence remain major challenges for CR. Khatanga et al used logistic regression to identify factors such as surgical diagnosis, non or former tobacco use and intensity of physician recommendation as independent predictors for CR participation, whereas factors including anxiety, depression, or executive function had no significant impact. Other studies have suggested that age, low socioeconomic status, gender, CR center location, psycho-behavioral factors including lack of motivation and reduced self-efficacy are barriers to participate in CR based on conventional statistics. Our review included two studies which approached the issue via ML methods. Filos et al. developed an ML based prediction model to predict intention to engage in CR, incorporating all the aforementioned factors as predictors in model training. This model achieved a high sensitivity of 0.945. Additionally, the ML algorithm identified the contribution of each risk factor to the ultimate outcomes, offering potential guidance for providers on prioritizing clinical interventions in cases where multiple risk factors coexist. Jahandideh et al predicted long term CR adherence using both identified risk factors mentioned above and physical fitness level, such as maximal quadriceps strength, which were not yet identified as independent risk factors. The model reached prediction accuracy at 71.5%. Some factors may be overlooked based on their individual statistical significance, but those factors may exert significant impact on ultimate outcome in combination.
with other factors. ML algorithms have the potential to identify complex interactions and patterns that traditional methods might miss\textsuperscript{56,57}.

**Conclusion**

There's a scarcity of ML-based clinical prognostic models for predicting healthcare outcomes in CR patients. While current models show good prediction capacities, their missing information and methodological biases make it challenging to determine the best model or rank their performance. Future research should focus on developing new prediction models aiming at various outcomes in more diverse populations with robust methodological approaches. Additionally, enhancing existing models' generalizability through external validation is needed. There's a long journey ahead before these models can be fully embraced in clinical settings.

**References**


12. Tessler J, Bordoni B. *Cardiac Rehabilitation*.; 2024.


